RESERVADO

INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO

UNIVERSIDADE TÉCNICA DE LISBOA

"BOOTSTRAP" ITERATIVO: APLICAÇÃO AO ÍNDICE DE GINI

Dissertação apresentada como requisito parcial para a obtenção do grau de mestre em Métodos Matemáticos para Economia e Gestão de Empresas

José Manuel Zorro Mendes

Junho 1990

A realização de um trabalho dificilmente pode ser considerada um acto puramente individual, já que, desde o assegurar condições básicas para a sua feitura, até à discussão de ideias, passando pela ajuda em certos aspectos menos dominados pelo Autor, são múltiplos os contactos estabelecidos.

O presente trabalho não é excepção e cabe aqui agradecer a quem muito o Autor deve.

Em primeiro lugar, ao orientador desta dissertação de Mestrado - o Professor Doutor Bento Murteira - com quem foram discutidas, não só as linhas gerais do trabalho, como também, inúmeros aspectos pontuais. De notar o seu empenhamento (e a sua paciência!) na revisão do texto final, no que consumiu muitas horas. Não menos importante, o entusiasmo e a confiança que soube transmitir ao Autor, ajudando a vencer momentos de desânimo que sempre aparecem.

Ao colega Dr. João M. Andrade e Silva, a quem se deve a elaboração do programa informático utilizado na parte empírica do trabalho (talvez a primeira pessoa, em Portugal, a implementar o "Bootstrap" informaticamente). Não só o Autor, mas, em geral, os adeptos do "Bootstrap", devem-lhe um sincero agradecimento.

À colega Dra. Dulce Ferreira, pelas muitas horas passadas a trabalhar com o processador de texto Word 3.01 da Macintosh, dando uma preciosa ajuda à obtenção da forma final do trabalho.

Finalmente, não poderia deixar de referir os meus pais que, desde sempre, se esforçaram para me dar todo o possível apoio moral e material ao prosseguimento dos estudos. Contei sempre e incondicionalmente com esse apoio para fazer este trabalho, que, por isso, também é, em certa medida, deles.

É óbvio que quaisquer erros ou omissões são da total responsabilidade do Autor, não respondendo por eles as pessoas agradecidas ou outras que tiverem colaborado com o Autor.

ÍNDICE

1 - INTRODUÇÃO

A aplicação da inferência estatística esbarra, não poucas vezes, com o completo desconhecimento da distribuição do universo em que se está a operar. Este desconhecimento prejudica a realização de inferências sobre o parâmetro (ou parâmetros) em estudo, o qual se supõe ser uma funcional da distribuição do universo, na medida em que não permite determinar a distribuição por amostragem do estimador, o que vem, consequentemente, dificultar a obtenção de importantes elementos, tais como: o valor esperado, o enviesamento, a variância e a construção de intervalos de confiança para o parâmetro.

Sucede outras vezes, que a distribuição do universo é conhecida, mas o parâmetro tem expressão funcional algo complexa, o que não permite determinar a distribuição do seu estimador. Nestas condições, acabam por ter-se as mesmas consequências negativas do caso anterior, ao nível da inferência estatística sobre o parâmetro em análise.

As duas situações anteriores, que se deparam com frequência no âmbito das aplicações práticas (nomeadamente, em estudos de natureza económica), constituem fonte de desânimo para o investigador e têm obstado à boa conclusão de muitos estudos empíricos.

Uma forma de ultrapassar os problemas atrás levantados, reside na aplicação do "Bootstrap", ou melhor, das técnicas "Bootstrap" (dada a diversidade de métodos existentes).

O objectivo do presente trabalho é, precisamente, o de apresentar os desenvolvimentos existentes quanto à aplicação das técnicas "Bootstrap" na construção de intervalos de confiança para um parâmetro unidimensional. Dá-se especial realçe ao caso não paramétrico (ou seja, quando a distribuição do universo é desconhecida), pois é nesta situação que a utilidade do "Bootstrap" se apresenta com mais intensidade.

Escolhem-se os intervalos de confiança como o indicador estatístico a tratar na aplicação dos métodos "Bootstrap", já que, quando se pretende inferir sobre um parâmetro desconhecido, um dos processos mais relevantes (senão o mais relevante, face ao desconhecimento do valor exacto do parâmetro) é o de construir um intervalo de valores que contenha o parâmetro em estudo, com determinado nível de confiança.

No ponto 2 do trabalho, apresenta-se, genericamente, o "Bootstrap", procurando-se relacioná-lo com outros métodos de reamostragem e explicitando-se algumas das suas aplicações. Por último, enuncia-se um resultado sobre o comportamento assintótico do "Bootstrap".

O ponto 3 ocupa-se da construção de intervalos de confiança através das técnicas "Bootstrap". Começam por apresentar-se os métodos expostos por Efron, os quais se encontram amplamente discutidos na literatura sobre o assunto, para depois se aprofundarem os métodos iterativos, menos difundidos, mas não menos interessantes, da autoria de Beran, Hall e Abramovitch e Singh (entre outros autores e com especial realçe para o primeiro).

Em relação ao ponto 3 do trabalho, há que esclarecer dois aspectos: o primeiro tem a ver com o facto de serem analisados apenas os métodos "Bootstrap" para construção de intervalos de confiança que podem chamar-se

de "puros", no sentido em que não conjugam o "Bootstrap" com as expansões de Edgeworth (apenas se faz uma referência breve a estes métodos no ponto 3.5 do trabalho); o segundo reside no facto de não se pretender fazer um "survey" de todos os métodos "Bootstrap" para a construção de intervalos de confiança - a literatura sobre o assunto é vastíssima e muitos autores limitam-se a introduzir pequenas alterações aos métodos propostos por Efron, dando origem a um sem número de metodologias que pouco acabam por diferir entre si e sobre as quais se torna difícil (senão mesmo impossível) elaborar juízos de valor comparativos.

No ponto 4, ensaia-se uma aplicação dos intervalos de confiança "Bootstrap" deduzidos teoricamente ao Índice de Gini para os rendimentos dos produtores agrícolas das regiões autónomas dos Açores e da Madeira. O Índice de Gini tem sido, desde há longos anos, um instrumental precioso na análise da desigualdade na distribuição dos rendimentos, mas cuja interpretação estatística tem permanecido na penumbra, em virtude de não se conhecer a distribuição por amostragem do seu estimador. Na verdade, não tem sentido dizer, por exemplo, que a concentração dos rendimentos na região A é superior à da região B, pelo simples facto de o Índice de Gini referente aos rendimentos, calculado com base em duas particulares amostras (uma para a região A, outra para a região B), ter dado o valor de 32% na região A e 30% na região B - os dois valores (calculados com base em amostras concretas, sublinhe-se), por si sós, não são suficientes para inferir com alguma credibilidade sobre os "verdadeiros" Índices de Gini, calculados sobre as duas populações. O "Bootstrap" permite a estimação da distribuição por amostragem do estimador do Índice de Gini, passando a comparação da concentração dos rendimentos entre as regiões A e B, do nível das amostras (sempre redutor e muito falível) para o nível da população.

Mais do que esgotar o tema, pretende-se realizar uma introdução ao "Bootstrap" e às suas técnicas de construção de intervalos de confiança não paramétricos, apresentando-o como um valioso método que permite ultrapassar situações de escassa informação. Os poucos anos de vida deste tema justificam a existência de lacunas na sua estrutura teórica, havendo aspectos menos consolidados e outros ainda longe de esclarecer. No entanto, a vitalidade da discussão teórica em torno do "Bootstrap" é por demais evidente, tudo levando a crer que as lacunas existentes serão preenchidas pouco a pouco, permitindo construir um edifício teórico que irá abrir mais uma das portas do conhecimento. Este trabalho atingirá os seus objectivos se conseguir mostrar que essa porta já se abriu um pouco.

2 - A IMPORTÂNCIA DO "BOOTSTRAP" NO CONTEXTO DA INFERÊNCIA ESTATÍSTICA

2.1 - APRESENTAÇÃO DO "BOOTSTRAP"

Considere-se um dado universo estatístico com função de distribuição F(x) e função de probabilidade ou de densidade de probabilidade f(x) (consoante a distribuição seja discreta ou contínua, respectivamente), ambas com domínio em $\mathbb{R}$.

O objectivo é obter informação sobre um parâmetro unidimensional θ, o qual depende da distribuição do universo e, consequentemente, da função de distribuição F(x),

$$\theta \equiv \theta[F(x)] \equiv \theta(F). \tag{2.1}$$

Para o efeito observa-se uma amostra casual, de dimensão n, do universo,

$$(X_1, X_2, ..., X_n), \quad X_i \overset{iid.}{\sim} F \ , i = 1, 2, ..., n, \tag{2.2}$$

onde F representa a distribuição do universo[1].

A amostra concreta é,

$$(X_1 = x_1, X_2 = x_2, ..., X_n = x_n).$$

---

[1] É costume, na literatura, representar a distribuição do universo pela correspondente função de distribuição $F(x) \equiv F$.
Assim, ao escrever-se $X_i \overset{iid.}{\sim} F, i = 1, 2, ..., n$, está-se a dizer que as variáveis aleatórias X_i são independentes e idênticamente distribuídas, de acordo com a função de distribuição F.

Para obter informação sobre o parâmetro θ, escolhe-se, em geral, um estimador, $\hat{\theta}$, o qual será função da amostra casual,

$$\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n). \tag{2.3}$$

É claro que a estimativa imediata de θ será,

$$\hat{\theta} \equiv \hat{\theta}(x_1, x_2, ..., x_n). \tag{2.4}$$

No estudo estatístico de um parâmetro, não interessa o simples conhecimento de uma estimativa. O grau de incerteza que caracteriza os problemas de inferência estatística motiva a construção de outros indicadores, que permitam aquilatar a qualidade estatística da variável aleatória utilizada como estimador do parâmetro. É neste âmbito, que se constrõem expressões para o valor esperado do estimador, para a variância do estimador ou para um intervalo de confiança de grau $(1-2\alpha)100\%$, $0 < \alpha < 0.5$, para o parâmetro. Para obter este tipo de informação, tem de conhecer-se a distribuição por amostragem do estimador $\hat{\theta}(X_1, X_2, ..., X_n)$. Como $\hat{\theta}$ depende de $X_1, X_2, ..., X_n$ e $X_i \overset{iid.}{\sim} F$, $i = 1, 2, ..., n$, por (2.2), a distribuição por amostragem de $\hat{\theta}$ depende da distribuição do universo,

$$\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n) \equiv \hat{\theta}(X_1, X_2, ..., X_n ; F).$$

Assim,

$$\mu \equiv E[\hat{\theta}] \equiv E_F[\hat{\theta}(X_1, X_2, ..., X_n)] \equiv \mu(F), \tag{2.5}$$

$$\sigma \equiv \sqrt{V(\hat{\theta})} \equiv \sqrt{V_F[\hat{\theta}(X_1, X_2, ..., X_n)]} \equiv \sigma(F), \tag{2.6}$$

$$P(a \leq \hat{\theta} \leq b) \equiv P_F[a \leq \hat{\theta}(X_1, X_2, ..., X_n) \leq b]. \tag{2.7}$$

Estão a considerar-se como parâmetros da distribuição do estimador $\hat{\theta}(X_1, X_2, ..., X_n)$, a média, o desvio padrão e os quantis. Podiam considerar-se quaisquer outros parâmetros, mas, a título exemplificativo, apenas se referem estes, por serem os mais notáveis e de maior uso.

É claro que, em regra, $\mu(F)$, $\sigma(F)$ e $P_F[a \leq \hat{\theta}(X_1, X_2, ..., X_n) \leq b]$, não dependem apenas da distribuição do universo, mas também, da forma do estimador $\hat{\theta}(X_1, X_2, ..., X_n)$ e da dimensão da amostra, n. No entanto, supondo que $\hat{\theta}(X_1, X_2, ..., X_n)$ e n são fixos, ou seja, que não está em causa a escolha de outro estimador para o parâmetro ou de outra dimensão para a amostra, a dependência situa-se apenas em relação à distribuição do universo.

O princípio "Bootstrap", extremamente simples, para estimar $\mu(F)$, $\sigma(F)$ e $P_F[a \leq \hat{\theta}(X_1, X_2, ..., X_n) \leq b]$, consiste em tornear a necessidade de se conhecer a distribuição do universo, F, substituindo-a pela distribuição empírica da amostra, $\hat{F}$. Assim, as estimativas "Bootstrap" de $\mu(F)$, $\sigma(F)$ e $P_F[a \leq \hat{\theta}(X_1, X_2, ..., X_n) \leq b]$ serão $\mu(\hat{F})$, $\sigma(\hat{F})$ e $P_{\hat{F}}[a \leq \hat{\theta}(X_1, X_2, ..., X_n) \leq b]$, respectivamente.

Como pode ver-se, a opção pelo "Bootstrap" adequa-se mais aos casos em que existe muito pouca informação sobre a distribuição do universo. Quanto maior for a informação sobre esta, menor será a utilidade do "Bootstrap", impondo-se outros métodos como mais aconselháveis. Na verdade, o grau de conhecimento que se tenha sobre a distribuição do universo é determinante, face ao caminho a seguir. Sintetizando, pode estar-se numa das seguintes situações:

1) Caso paramétrico - conhece-se a distribuição do universo (logo, conhece-se F) e sabe-se que ela é caracterizada por determinados parâmetros, os quais são desconhecidos.

1.1) Consegue deduzir-se, teoricamente, a distribuição do estimador $\hat{\theta}(X_1, X_2, ..., X_n)$, a partir da distribuição do universo.

Sabendo a distribuição do estimador $\hat{\theta}(X_1, X_2, ..., X_n)$ (exacta ou assintótica), podem fazer-se inferências sobre os parâmetros $\mu(F)$ e $\sigma(F)$, ou sobre as probabilidades $P_F[a \leq \hat{\theta}(X_1, X_2, ..., X_n) \leq b]$.

Neste caso, pode obter-se informação relevante sobre o parâmetro θ, sem necessidade de recorrer ao "Bootstrap" ou a quaisquer outras técnicas similares.

1.2) Não se consegue deduzir, teoricamente, a distribuição do estimador $\hat{\theta}(X_1, X_2, ..., X_n)$, a partir da distribuição do universo.

Neste caso, não se conseguem fazer inferências sobre os parâmetros de $\hat{\theta}(X_1, X_2, ..., X_n)$ - $\mu(F)$, $\sigma(F)$ - ou sobre as probabilidades $P_F[a \leq \hat{\theta}(X_1, X_2, ..., X_n) \leq b]$.

A solução "Bootstrap" para este caso é gerar, aleatoriamente, B amostras (B suficientemente grande) de dimensão n, a partir da distribuição do universo (após ter estimado os parâmetros) e, para cada uma dessas amostras, calcular o valor concreto de $\hat{\theta}(X_1, X_2, ..., X_n)$,

$\hat{\theta}^{*1} \equiv \hat{\theta}(x^{*1}_1, x^{*1}_2, ..., x^{*1}_n)$, onde $(x^{*1}_1, x^{*1}_2, ..., x^{*1}_n)$, é a primeira amostra gerada aleatoriamente, a partir da distribuição estimada do universo.

$\hat{\theta}^{*2} \equiv \hat{\theta}(x^{*2}_1, x^{*2}_2, ..., x^{*2}_n)$, onde $(x^{*2}_1, x^{*2}_2, ..., x^{*2}_n)$, é a segunda amostra gerada aleatoriamente, a partir da distribuição estimada do universo.

...

$\hat{\theta}^{*B} \equiv \hat{\theta}(x^{*B}_1, x^{*B}_2, ..., x^{*B}_n)$, onde $(x^{*B}_1, x^{*B}_2, ..., x^{*B}_n)$, é a B-ésima amostra gerada aleatoriamente, a partir da distribuição estimada do universo.

A distribuição empírica dos valores $\hat{\theta}^{*1}$, $\hat{\theta}^{*2}$, ..., $\hat{\theta}^{*B}$ aproxima a distribuição do estimador $\hat{\theta}(X_1, X_2, ..., X_n)$, podendo estimar-se os parâmetros e as probabilidades já referidos da seguinte forma,

$$\hat{\mu}_{BOOT} = (1/B) \sum_{i=1}^{B} \hat{\theta}^{*i}, \tag{2.8}$$

$$\hat{\sigma}_{BOOT} = \sqrt{[1/(B-1)] \sum_{i=1}^{B} (\hat{\theta}^{*i} - \hat{\mu}_{BOOT})^2}, \tag{2.9}$$

$$\hat{P}_{BOOT}(a \le \hat{\theta} \le b) = \frac{\#\{a \le \hat{\theta}^{*i} \le b\}}{B}, \tag{2.10}$$

onde $\hat{\mu}_{BOOT}$, $\hat{\sigma}_{BOOT}$ e $\hat{P}_{BOOT}(a \leq \hat{\theta} \leq b)$ são as estimativas "Bootstrap" de $\mu(F)$, $\sigma(F)$ e $P_F[a \leq \hat{\theta}(X_1, X_2, ..., X_n) \leq b]$, respectivamente.

Repare-se que, quando $B \to +\infty$, tem-se,

$$\hat{\mu}_{BOOT} = \mu(F_{\hat{\eta}}) \equiv E_{F_{\hat{\eta}}}[\hat{\theta}(X_1, X_2, ..., X_n)], \tag{2.11}$$

$$\hat{\sigma}_{BOOT} = \sigma(F_{\hat{\eta}}) \equiv \sqrt{V_{F_{\hat{\eta}}}[\hat{\theta}(X_1, X_2, ..., X_n)]}, \tag{2.12}$$

$$\hat{P}_{BOOT}(a \leq \hat{\theta} \leq b) = P_{F_{\hat{\eta}}}[a \leq \hat{\theta}(X_1, X_2, ..., X_n) \leq b]. \tag{2.13}$$

Neste contexto, $F_{\hat{\eta}}$ é a função de distribuição que corresponde à distribuição do universo estimada (recorde-se que se está no caso em que a distribuição do universo é conhecida, mas os seus parâmetros são desconhecidos, tendo de ser estimados, o que dá origem a uma distribuição do universo estimada). Assim, se $F \equiv F(x \mid \eta_1, \eta_2, ..., \eta_q)$, ou seja, se a distribuição do universo for caracterizada por q parâmetros, $\eta_1, \eta_2, ..., \eta_q$, cujas estimativas são $\hat{\eta}_1, \hat{\eta}_2, ..., \hat{\eta}_q$, tem-se,

$$F_{\hat{\eta}} \equiv F(x \mid \hat{\eta}_1, \hat{\eta}_2, ..., \hat{\eta}_q). \tag{2.14}$$

Em rigor, as estimativas "Bootstrap" de $\mu(F)$, $\sigma(F)$ e $P_F[a \leq \hat{\theta}(X_1, X_2, ..., X_n) \leq b]$ são $\mu(F_{\hat{\eta}})$, $\sigma(F_{\hat{\eta}})$ e $P_{F_{\hat{\eta}}}[a \leq \hat{\theta}(X_1, X_2, ..., X_n) \leq b]$, respectivamente - o princípio "Bootstrap" está em substituir F por $F_{\hat{\eta}}$. No entanto, assim como não se consegue deduzir a distribuição do estimador $\hat{\theta}(X_1, X_2, ..., X_n)$ a partir da distribuição do universo, também não se consegue deduzir a sua distribuição a partir da distribuição do universo estimada - daí que se utilize o método de Monte Carlo, baseado na extracção de sucessivas amostras de um universo com a distribuição estimada, o qual dá origem às estimativas $\hat{\mu}_{BOOT}$, $\hat{\sigma}_{BOOT}$ e $\hat{P}_{BOOT}(a \leq \hat{\theta} \leq b)$. Na linguagem corrente, estas estimativas também são chamadas de "estimativas Bootstrap", embora haja

aqui uma incorrecção, dado que são apenas uma aproximação às "verdadeiras" estimativas "Bootstrap" e só as igualam quando $B \to +\infty$.

Para evitar linguagem mais pesada, vai passar a designar-se $\mu(F_{\hat{\eta}})$, $\sigma(F_{\hat{\eta}})$, $P_{F_{\hat{\eta}}}[a \le \hat{\theta}(X_1, X_2, ..., X_n) \le b]$, $\hat{\mu}_{BOOT}$, $\hat{\sigma}_{BOOT}$ e $\hat{P}_{BOOT}(a \le \hat{\theta} \le b)$ indistintamente como estimativas "Bootstrap", se bem que se tenha sempre presente a diferença entre os dois conjuntos de estimativas.

O método que tem vindo a expor-se denomina-se "Bootstrap paramétrico", precisamente por a distribuição do universo não ser totalmente desconhecida, faltando apenas o conhecimento de alguns parâmetros, os quais são previamente estimados, antes de se iniciar a aplicação do "Bootstrap" propriamente dito.

2) Caso não paramétrico - não se conhece a distribuição do universo (logo, desconhece-se F), assim como se ignora a existência de eventuais parâmetros que a caracterizem.

Está-se num caso de desconhecimento quase total, em que não se conseguem fazer as habituais inferências estatísticas sobre os parâmetros de $\hat{\theta}(X_1, X_2, ..., X_n)$ - $\mu(F)$, $\sigma(F)$ - ou sobre as probabilidades $P_F[a \le \hat{\theta}(X_1, X_2, ..., X_n) \le b]$.

A solução "Bootstrap" para este caso é gerar, aleatoriamente, B amostras (B suficientemente grande) de dimensão n, a partir da particular amostra observada $(x_1, x_2, ..., x_n)$ (aqui está a grande diferença em relação ao "Bootstrap paramétrico", devido ao desconhecimento da distribuição do universo) - trata-

-se de efectuar tiragens com reposição a partir de $(x_1, x_2, ..., x_n)$. Para cada uma dessas amostras, calcula-se o valor concreto de $\hat{\theta}(X_1, X_2, ..., X_n)$ - $\hat{\theta}^{*1}$, $\hat{\theta}^{*2}$, ..., $\hat{\theta}^{*B}$ - tal como foi explicitado no caso 1.2.

A distribuição empírica dos valores $\hat{\theta}^{*1}$, $\hat{\theta}^{*2}$, ..., $\hat{\theta}^{*B}$, aproxima a distribuição do estimador $\hat{\theta}(X_1, X_2, ..., X_n)$, podendo-se estimar parâmetros e probabilidades, tal como foi indicado em (2.8), (2.9) e (2.10).

Mantêm-se válidas as igualdades (2.11), (2.12) e (2.13), quando $B \to +\infty$, substituindo $F_{\hat{\eta}}$ por $\hat{F}$,

$$\hat{\mu}_{BOOT} = \mu(\hat{F}) \equiv E_{\hat{F}}[\hat{\theta}(X_1, X_2, ..., X_n)], \tag{2.15}$$

$$\hat{\sigma}_{BOOT} = \sigma(\hat{F}) \equiv \sqrt{V_{\hat{F}}[\hat{\theta}(X_1, X_2, ..., X_n)]}, \tag{2.16}$$

$$\hat{P}_{BOOT}(a \le \hat{\theta} \le b) \equiv P_{\hat{F}}[a \le \hat{\theta}(X_1, X_2, ..., X_n) \le b]. \tag{2.17}$$

Também se mantêm válidas as considerações feitas (no caso 1.2) àcerca da distinção entre $\mu(F_{\hat{\eta}})$, $\sigma(F_{\hat{\eta}})$ e $P_{F_{\hat{\eta}}}[a \le \hat{\theta}(X_1, X_2, ..., X_n) \le b]$, por um lado, e $\hat{\mu}_{BOOT}$, $\hat{\sigma}_{BOOT}$ e $\hat{P}_{BOOT}(a \le \hat{\theta} \le b)$, por outro, as quais são agora transpostas para a distinção entre $\mu(\hat{F})$, $\sigma(\hat{F})$ e $P_{\hat{F}}[a \le \hat{\theta}(X_1, X_2, ..., X_n) \le b]$, por um lado, e $\hat{\mu}_{BOOT}$, $\hat{\sigma}_{BOOT}$ e $\hat{P}_{BOOT}(a \le \hat{\theta} \le b)$, por outro: em rigor, as estimativas "Bootstrap" de $\mu(F)$, $\sigma(F)$ e $P_F[a \le \hat{\theta}(X_1, X_2, ..., X_n) \le b]$ são aquelas que se obtêm substituindo F por $\hat{F}$, ou seja, $\mu(\hat{F})$, $\sigma(\hat{F})$ e $P_{\hat{F}}[a \le \hat{\theta}(X_1, X_2, ..., X_n) \le b]$, respectivamente, e as estimativas $\hat{\mu}_{BOOT}$, $\hat{\sigma}_{BOOT}$ e $\hat{P}_{BOOT}(a \le \hat{\theta} \le b)$ apenas coincidem com as "verdadeiras" estimativas "Bootstrap", quando $B \to +\infty$.

A diferença em relação ao caso 1.2 está na substituição da função de distribuição $F_{\hat{\eta}}$ pela função de distribuição $\hat{F}$. Neste contexto, $\hat{F}$ é a função de distribuição empírica da amostra,

$$\hat{F} \equiv \hat{F}(x) = \frac{\#\{x_i \leq x\}}{n}, \quad -\infty < x < +\infty. \tag{2.18}$$

Recorde-se que a distribuição empírica da amostra é aquela que atribui probabilidade 1/n a cada um dos n valores da amostra $(x_1, x_2, ..., x_n)$. Ora, quando se efectuam tiragens com reposição da amostra observada para gerar as amostras "Bootstrap" $(x_1^{*b}, x_2^{*b}, ..., x_n^{*b})$, b = 1, 2, ..., B, cada valor x_i, i = 1, 2, ..., n, tem igual probabilidade 1/n de ser escolhido, em cada tiragem.

Em conclusão, pode dizer-se que a distribuição empírica da amostra, $\hat{F}$, é a estimativa "Bootstrap" da distribuição do universo (aliás, $\hat{F}$ é a estimativa da máxima verosimilhança não paramétrica da distribuição do universo, F [veja-se Efron e Tibshirani (1986) - pg. 56]).

A aproximação da distribuição do universo, totalmente desconhecida, pela distribuição empírica da amostra, perfeitamente conhecida, é a base fundamental do método "Bootstrap" no caso não paramétrico - no domínio não paramétrico, está-se em presença do "Bootstrap genuíno", o qual, como foi dito, também pode ser aplicado em domínios paramétricos, embora aí, a existência de mais informação sobre a distribuição do universo, leve a aplicar o "Bootstrap paramétrico" (caso 1.2) ou a clássica análise teórica (caso 1.1).

Não é de mais realçar que o "Bootstrap", entendido nesta sua versão genuína, ou pura, passa de uma estrutura totalmente desconhecida, retratada por (2.2), para uma estrutura conhecida, caracterizada por,

$$(X_1^*, X_2^*, ..., X_n^*), \quad X_i^* \overset{iid.}{\sim} \hat{F}, \; i = 1, 2, ..., n, \tag{2.19}$$

onde X_i^*, i = 1, 2, ..., n, são as variáveis aleatórias "Bootstrap", cada uma das quais pode assumir qualquer valor de entre os valores da amostra observada $(x_1, x_2, ..., x_n)$. A função de probabilidade de cada X_i^* é dada por,

$$P(X_i^* = x) = \begin{cases} 1/n, & x = x_j, \quad j = 1, 2, \ldots, n \\ 0, & x \neq x_j, \quad j = 1, 2, \ldots, n \end{cases},$$

$i = 1, 2, \ldots, n.$

As igualdades (2.11), (2.12) e (2.13), apresentadas no âmbito do "Bootstrap paramétrico", e as igualdades (2.15), (2.16) e (2.17), apresentadas no âmbito do "Bootstrap genuíno" (que passa a designar-se simplesmente por "Bootstrap"), só são válidas, quando $B \to +\infty$, se a dimensão das amostras "Bootstrap" for n, ou seja, se for igual à dimensão da amostra observada [veja-se Efron e Tibshirani (1986) - pg. 56].

Como pode ver-se, o "Bootstrap" apresenta um meio expedito de fazer inferências em múltiplas situações, cuja valia e competitividade, face a métodos mais tradicionais, se fazem sentir, com especial acuidade, nos casos em que a ausência de informação sobre o universo estatístico em causa é quase total, circunscrevendo-se, praticamente, à particular amostra observada. Na verdade, nos problemas não paramétricos, o "Bootstrap" permite resolver situações de impasse, ao proporcionar a obtenção de uma gama completa de informações sobre o parâmetro em estudo, exclusivamente a partir da amostra observada. É neste aspecto que reside a sua importância nos domínios da inferência estatística - o "Bootstrap" permite ultrapassar os casos em que o desconhecimento sobre a distribuição do universo impera (em maior ou menor grau), com uma certa elegância e com a validade estatística que, mais à frente, se analisa.

Nos pontos seguintes, o estudo vai situar-se ao nível do caso 2 - caso não paramétrico - salvo referência explícita em contrário.

2.2 - O "BOOTSTRAP" COMO UM MÉTODO DE REAMOSTRAGEM

As características inovadoras do "Bootstrap" não devem levar a concluir que se trata de um caso isolado, sem nenhuma relação com outros métodos ou famílias de métodos existentes. Na verdade, o "Bootstrap" insere-se numa família - os chamados métodos de reamostragem - que procura ultrapassar o problema da escassez de informação àcerca do universo, ao realizar inferências sobre um dado parâmetro a partir do reprocessamento da única informação conhecida: a particular amostra observada.

Os métodos de reamostragem são assim chamados, porque constroem novas amostras, a partir da amostra observada, baseando-se essa construção, num esquema empírico de probabilidades atribuído aos valores da amostra observada.

Suponha-se a situação definida por (2.1) e (2.2), em que se querem realizar inferências sobre um parâmetro, $\theta \equiv \theta(F)$, funcional da distribuição do universo, com base numa amostra casual de dimensão n, $(X_1, X_2, ..., X_n)$, $X_i \overset{iid}{\sim} F$, $i = 1, 2, ..., n$, cuja observação deu os valores $(x_1, x_2, ..., x_n)$.

Sendo F desconhecida, a única informação disponível é a concreta amostra observada $(x_1, x_2, ..., x_n)$, daí que a filosofia dos métodos de reamostragem seja a de substituir uma estrutura desconhecida, definida por (2.2), por uma estrutura conhecida, caracterizada por,

$$(X_1^*, X_2^*, ..., X_n^*), \quad X_i^* \overset{iid.}{\sim} \tilde{F} \ , i = 1, 2, ..., n, \tag{2.20}$$

onde X_i^*, i = 1, 2, ..., n são variáveis aleatórias, cada uma das quais pode assumir valores de entre os valores da amostra observada $(x_1, x_2, ..., x_n)$, de acordo com o esquema empírico de probabilidades atribuído aos valores x_i, i = 1, 2, ..., n, o qual é aqui representado por $\tilde{F}$. A função de probabilidade de cada X_i^* é dada por,

$$P(X_i^* = x) = \begin{cases} \tilde{P}_j, & x = x_j, \quad j = 1, 2, ..., n \\ 0, & x \neq x_j, \quad j = 1, 2, ..., n \end{cases},$$

$$i=1, 2, ..., n,$$

onde $\tilde{P}_j$ é a probabilidade atribuída a x_j, j = 1, 2, ..., n. Naturalmente, os $\tilde{P}_j$ obedecem às seguintes condições,

$$\sum_{j=1}^{n} \tilde{P}_j = 1 \ ; \ \tilde{P}_j \geq 0, j = 1, 2, ..., n.$$

A aplicação dos métodos de reamostragem passa pelo cálculo dos valores $\hat{\theta}^{*b}$, b = 1, 2, ..., B (relembre-se que $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$ é o estimador escolhido de θ), o que se efectua da seguinte forma,

$$\hat{\theta}^{*b} \equiv \hat{\theta}(x_1^{*b}, x_2^{*b}, ..., x_n^{*b}), b = 1, 2, ..., B, \tag{2.21}$$

onde $(x_1^{*b}, x_2^{*b}, ..., x_n^{*b})$, é a b-ésima amostra gerada aleatoriamente, a partir da amostra observada $(x_1, x_2, ..., x_n)$ e segundo o esquema empírico de probabilidades definido por $\tilde{F}$.

A distribuição empírica dos valores $\hat{\theta}^{*b}$, b = 1, 2, ..., B, aproxima a distribuição do estimador $\hat{\theta}(X_1, X_2, ..., X_n)$, podendo-se realizar inferências sobre o parâmetro θ.

Repare-se que, se, em (2.20), se substituir X_i° por X_i^*, i = 1, 2, ..., n e $\tilde{F}$ por $\hat{F}$, tem-se o "Bootstrap". Neste caso, $\tilde{P}_j = 1/n = P_j^\circ$, j = 1, 2, ..., n, porque a distribuição empírica da amostra, $\hat{F}$, atribui probabilidade 1/n a cada valor da amostra observada $(x_1, x_2, ..., x_n)$. Em vez das amostras $(x_1^{\circ b}, x_2^{\circ b}, ..., x_n^{\circ b})$, b = 1, 2, ..., B, passam a valer as amostras "Bootstrap" $(x_1^{*b}, x_2^{*b}, ..., x_n^{*b})$, b = 1, 2, ..., B, e, em vez de $\hat{\theta}^{\circ b}$, b = 1, 2, ..., B, têm-se os valores "Bootstrap" do estimador $\hat{\theta}(X_1, X_2, ..., X_n)$, $\hat{\theta}^{*b}$, b = 1, 2, ..., B.

Convém aproveitar esta ocasião, para fazer uma clara distinção entre $\hat{F}$ e o que na literatura sobre o "Bootstrap" se designa por $\hat{F}^*$. Por $\hat{F}$ entende-se a distribuição empírica da amostra observada $(x_1, x_2, ..., x_n)$, a qual atribui probabilidade $P_j^\circ = 1/n$ a cada valor x_j, j = 1, 2, ..., n. Por $\hat{F}^*$ entende-se a distribuição empírica da amostra "Bootstrap" observada $(x_1^*, x_2^*, ..., x_n^*)$, a qual atribui probabilidade,

$$P_j^* = \frac{\#\{x_i^* = x_j\}}{n}, \tag{2.22}$$

a cada valor x_j, j = 1, 2, ..., n.

Se, em (2.22), substituirmos x_i^* por X_i^*, ou seja, se, em vez de considerarmos uma amostra "Bootstrap" concreta $(x_1^*, x_2^*, ..., x_n^*)$, tivermos em conta uma amostra "Bootstrap" genérica aleatória, $(X_1^*, X_2^*, ..., X_n^*)$, o vector $P^* \equiv (P_1^*, P_2^*, ..., P_n^*)$, chamado de vector de reamostragem, pode ser entendido como um vector aleatório, com distribuição multinomial, reescalada pelo factor 1/n,

$$P^* \underset{*}{\sim} \frac{\text{Mult}_n(n, P^\circ)}{n}, \tag{2.23}$$

onde,

$$\text{Mult}_n(n,P^\circ) = \frac{n!}{N_1^*!\ N_2^*!\ \ldots\ N_n^*!}\ (1/n)^{N_1^*}\ (1/n)^{N_2^*} \ldots (1/n)^{N_n^*}\ .$$

Tenha-se em atenção que:

- O simbolo $\overset{*}{\sim}$ significa que a distribuição do vector aleatório P^* está dependente do universo de amostras "Bootstrap" $(X_1^*, X_2^*, \ldots, X_n^*)$ que são possíveis de construir, a partir da particular amostra $(x_1, x_2, \ldots, x_n)$.

- O simbolo P° designa outro vector de reamostragem - precisamente o resultante da distribuição empírica da amostra observada, $P^\circ \equiv (P_1^\circ, P_2^\circ, \ldots, P_n^\circ) = (1/n, 1/n, \ldots, 1/n)$.

- O simbolo N_j^*, $j = 1, 2, \ldots, n$, é definido como, $N_j^* = \#\{x_i^* = x_j\} = n\ P_j^*$.

Daqui vem que $N^* = (N_1^*, N_2^*, \ldots, N_n^*) \overset{*}{\sim} \text{Mult}_n(n,P^\circ)$.

Outro método de reamostragem, muito conhecido e mais antigo do que o "Bootstrap", é o "Jackknife".

O "Jackknife", a partir da concreta amostra observada $(x_1, x_2, \ldots, x_n)$, constrói n amostras, cada uma delas de dimensão n-1, por se ter excluído o valor x_i, $i = 1, 2, \ldots, n$. Assim, a primeira amostra será $(x_2, x_3, \ldots, x_n)$, a segunda será $(x_1, x_3, \ldots, x_n)$, e assim sucessivamente, até à n-ésima amostra, $(x_1, x_2, \ldots, x_{n-1})$.

Para ver no "Jackknife" um particular método de reamostragem, substitua-se (2.20) por,

$$(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n),\ X_j \overset{iid}{\sim} F_{(i)},\ j = 1, 2, \ldots, i-1, i+1, \ldots, n;$$

$$i = 1, 2, \ldots, n, \qquad (2.24)$$

onde $F_{(i)}$ representa a distribuição empírica da amostra concreta observada $(x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n)$, atribuindo probabilidade $1/(n-1)$ a cada valor x_j, $j = 1, 2, ..., i-1, i+1, ..., n$ e probabilidade zero a x_i. Neste caso, $\tilde{P}_j = 1/(n-1)$, $j = 1, 2, ..., i-1, i+1, ..., n$ e $\tilde{P}_i = 0$, o que dá o seguinte vector de reamostragem,

$$P_{(i)} = \left(\frac{1}{n-1}, \frac{1}{n-1}, \cdots, \frac{1}{n-1}, 0, \frac{1}{n-1}, \cdots, \frac{1}{n-1}\right),$$

onde o zero está na i-ésima posição. Em vez das amostras $(x_1^{*b}, x_2^{*b}, ..., x_n^{*b})$, $b = 1, 2, ..., B$, passam a valer as amostras "Jackknife" $(x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n)$, $i = 1, 2, ..., n$, e, em vez de $\hat{\theta}^{*b}$, $b = 1, 2, ..., B$, têm-se os valores "Jackknife" do estimador $\hat{\theta}(X_1, X_2, ..., X_n)$, $\hat{\theta}_{(i)} \equiv \hat{\theta}(x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n)$, $i = 1, 2, ..., n$.

Veja-se, agora, a relação existente entre o "Jackknife" e o "Bootstrap".

Considere-se a estatística (2.31), $R \equiv R(X_1, X_2, ..., X_n ; F)$, que adiante se retoma. Defina-se,

$$\hat{R} \equiv R(P^\circ) \equiv R(x_1, x_2, ..., x_n ; \hat{F}), \tag{2.25}$$

como o valor da estatística R, após se ter observado $(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$ e,

$$R^* \equiv R(P^*) \equiv R(X_1^*, X_2^*, ..., X_n^* ; \hat{F}), \tag{2.26}$$

como a estatística R, sendo função, não das variáveis aleatórias originais do universo, mas sim das variáveis aleatórias "Bootstrap". Repare-se que, se forem construídas B amostras (B suficientemente grande) "Bootstrap", a distribuição empírica dos valores $R^{*b} \equiv R(x_1^{*b}, x_2^{*b}, ..., x_n^{*b} ; \hat{F})$, $b = 1, 2, ..., B$, aproxima a distribuição de R^*; por outro lado, como R^* é a aproximação "Bootstrap" de R, pode dizer-se que a distribuição empírica dos valores R^{*b}, $b = 1, 2, ..., B$, aproxima a distribuição de R.

Então, no "Bootstrap", $R^* \equiv R(P^*)$ é utilizada para realizar inferências sobre R. Em vez de $R^* \equiv R(P^*)$, vai agora considerar-se uma aproximação, dada pela expansão em fórmula de Taylor, até à terceira ordem, no ponto P°,

$$R^* \equiv R(P^*) \cong R(P^\circ) + (P^* - P^\circ)^T U + (1/2)\,(P^* - P^\circ)^T V\,(P^* - P^\circ), \qquad (2.27)$$

onde,

$$P^\circ = \left[\frac{1}{n}\ \frac{1}{n}\ \cdots\ \frac{1}{n}\right]^T,$$

$$P^* = \left[P_1^*\ P_2^*\ \cdots\ P_n^*\right]^T,$$

$$U = \left[\frac{\partial R(P^*)}{\partial P_1^*}\ \frac{\partial R(P^*)}{\partial P_2^*}\ \cdots\ \frac{\partial R(P^*)}{\partial P_n^*}\right]^T_{P^* = P^\circ},$$

$$V = \left[\begin{array}{cccc} \dfrac{\partial^2 R(P^*)}{\partial P_1^{*2}} & \dfrac{\partial^2 R(P^*)}{\partial P_1^* \partial P_2^*} & \cdots & \dfrac{\partial^2 R(P^*)}{\partial P_1^* \partial P_n^*} \\ \dfrac{\partial^2 R(P^*)}{\partial P_2^* \partial P_1^*} & \dfrac{\partial^2 R(P^*)}{\partial P_2^{*2}} & \cdots & \dfrac{\partial^2 R(P^*)}{\partial P_2^* \partial P_n^*} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\partial^2 R(P^*)}{\partial P_n^* \partial P_1^*} & \dfrac{\partial^2 R(P^*)}{\partial P_n^* \partial P_2^*} & \cdots & \dfrac{\partial^2 R(P^*)}{\partial P_n^{*2}} \end{array}\right]_{P^* = P^\circ}.$$

Repare-se que:

- anteriormente tinha-se definido P° e P^* como énuplos ordenados - $(1/n, 1/n, ..., 1/n)$ e $(P_1^*, P_2^*, ..., P_n^*)$, respectivamente - agora interessa defini-los como matrizes coluna com n linhas;

- a matriz U é uma matriz coluna com as derivadas parciais de (2.26), tomadas no ponto $P^* = P^\circ$;

- a matriz V é uma matriz quadrada, nxn, com as derivadas parciais de segunda ordem de (2.26), tomadas no ponto $P^* = P^\circ$.

Calcule-se $E_*[R(P^*)]$ [2], através de (2.27),

$$E_*[R(P^*)] \cong R(P^\circ) + [E_*(P^*) - P^\circ]^T U + (1/2)\, E_*[(P^* - P^\circ)^T V (P^* - P^\circ)].$$

Como $E_*(P^*) = P^\circ$ e $E_*[(P^* - P^\circ)^T V (P^* - P^\circ)] = tr\{V[(1/n^2) I - (1/n^3) e^T e]\}$ [veja-se Efron (1979) - pg. 13], onde I é a matriz identidade de ordem n e a letra "e" representa uma matriz linha com n colunas, com todos os elementos iguais a 1, ou seja, $e = [1\ 1\ ...\ 1]$, tem-se,

$$E_*[R(P^*)] \cong R(P^\circ) + (1/2)\, tr\{V[(1/n^2) I - (1/n^3) e^T e]\} =$$

$$= R(P^\circ) + (1/2n)\, \bar{V}, \qquad (2.28)$$

onde $\bar{V} = \sum_{i=1}^{n} (1/n)\ V_{ii}$, V_{ii} - elemento ii da matriz V [veja-se Efron (1979) - pg. 13].

Seguidamente considere-se a aproximação de $R^* \equiv R(P^*)$ dada pela expansão em fórmula de Taylor, até à segunda ordem, no ponto P°,

$$R^* \equiv R(P^*) \cong R(P^\circ) + (P^* - P^\circ)^T U, \qquad (2.29)$$

---

[2] O índice * significa que este valor esperado é calculado em relação ao universo de amostras "Bootstrap" $(X_1^*, X_2^*, ..., X_n^*)$ que são passíveis de construir, a partir da particular amostra $(x_1, x_2, ..., x_n)$.

e calcule-se $V_*[R(P^*)]$ [3], através de (2.29),

$$V_*[R(P^*)] \cong U^T[(1/n^2)\, I - (1/n^3)\, e^T e]\, U = \sum_{i=1}^{n} (U_i^2/n^2)\,, \qquad (2.30)$$

onde U_i é o i-ésimo elemento da matriz U [veja-se Efron (1979) - pg. 13].

Os resultados (2.28) e (2.30) são, basicamente, as estimativas "Jackknife" para o valor esperado e a variância da estatística $R \equiv R(X_1, X_2, ..., X_n ; F)$ [veja-se Efron (1979) - pg. 13] [4] - daí que se possa concluir que o "Jackknife" é, também, um método "Bootstrap", embora não aplicado à estatística $R \equiv R(X_1, X_2, ..., X_n ; F)$, mas sim a uma aproximação (que pode ser linear, como no caso da estimativa da variância acima apresentado [veja-se Efron (1982a) - pg. 39]), dada pela expansão em fórmula de Taylor.

Para terminar este ponto, sobre os métodos de reamostragem, veja-se outra forma, particularmente interessante, de atribuir um vector de

---

[3] O índice * tem o mesmo significado do que o explicitado para o valor esperado $E_*[R(P^*)]$.

[4] Reforce-se que as expressões (2.28) e (2.30) não são exactamente as estimativas "Jackknife" de $E_F(R)$ e $V_F(R)$, respectivamente, já que, existem algumas diferenças entre as referidas expressões e as verdadeiras estimativas "Jackknife". Centrando-nos no caso da variância (por exemplo), vão explicitar-se essas diferenças.

Em primeiro lugar, as estimativas "Jackknife" substituem as derivadas $U_i = \dfrac{\partial R(P^*)}{\partial P_i^*}$, pelas diferenças finitas $\tilde{U}_i = (n-1)[R_\bullet - R_{(i)}]$, onde $R_{(i)} = R(x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n)$ e $R_\bullet = \sum_{i=1}^{n} (1/n)\, R_{(i)}$.

Em segundo lugar, a expressão (2.30) aparece multiplicada pelo factor $n/(n-1)$.

Assim, a "verdadeira" estimativa "Jackknife" de $V_F(R)$ é $\hat{\sigma}^2_{JACK} = \sum_{i=1}^{n} [1/n(n-1)]\, \tilde{U}_i^2$, a qual não coincide exactamente com (2.30).

No entanto, estas pequenas diferenças entre as expressões (2.28) e (2.30) e as estimativas "Jackknife" não são suficientes para invalidar as conclusões que se vão tirar.

reamostragem à concreta amostra observada, dando origem ao chamado "Bootstrap Bayesiano".

Considere-se o seguinte algoritmo:

1 - A partir de uma variável aleatória uniforme contínua, no intervalo [0 ; 1], gerem-se, aleatoriamente, n-1 valores: $u_1, u_2, ..., u_{n-1}$.

2 - Faça-se $u_{(0)} = 0$ e $u_{(n)} = 1$ e calculem-se as diferenças $g_i = u_{(i)} - u_{(i-1)}$, $i = 1, 2, ..., n$. Note-se que $u_{(1)} < u_{(2)} < ... < u_{(n-1)}$, representa os u_i, $i = 1, 2, ..., n-1$, dispostos por ordem crescente.

3 - Após ter-se concluído o passo anterior, o vector de reamostragem atribuído à amostra observada $(x_1, x_2, ..., x_n)$ é $(g_1, g_2, ..., g_n)$, ou seja, g_i é a probabilidade atribuída a x_i, $i = 1, 2, ..., n$.

Considerando bastantes reamostragens, isto é, repetindo B vezes (B suficientemente grande) os passos 1 e 2 do algoritmo, chega-se a um vector de reamostragem $G \equiv (G_1, G_2, ..., G_n)$, para a particular amostra observada $(x_1, x_2, ..., x_n)$. Com base neste vector G, pode aproximar-se a distribuição do estimador $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$ e realizar inferências sobre o parâmetro θ.

Este método é chamado de "Bootstrap Bayesiano", porque cada réplica "Bootstrap", dada pelo algoritmo acima expresso, gera uma probabilidade à posteriori, g_i, para cada x_i, $i = 1, 2, ..., n$ [veja-se Rubin (1981) - pg. 131]. Os valores de X (variável aleatória genérica do universo) não observados, isto é, que não fazem parte da amostra $(x_1, x_2, ..., x_n)$, têm probabilidade à posteriori nula. As probabilidades à posteriori estão centradas em 1/n, $E(G_i) = 1/n$, $i = 1, 2, ..., n$, ou seja, as probabilidades Bayesianas estão centradas nas probabilidades empíricas da amostra observada $(x_1, x_2, ..., x_n)$. [veja-se Rubin (1981) - pg. 131].

Como acaba de ver-se, o "Bootstrap", ou variantes suas, apresenta grandes afinidades com os processos de reamostragem, devendo inserir-se nesta vasta e promissora família de métodos.

2.3 - ALGUMAS APLICAÇÕES DO "BOOTSTRAP"

No ponto 2.1 deste trabalho apresentou-se o "Bootstrap", como um método que possibilita a realização de inferências sobre um parâmetro funcional da distribuição do universo, $\theta \equiv \theta(F)$, através da construção da distribuição "Bootstrap" de um seu estimador, $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$.

No entanto, a contribuição do "Bootstrap" para a inferência estatística pode ser alargada a outros domínios, o que vai exemplificar-se sem a preocupação de apresentar todas as aplicações do "Bootstrap" (o que seria tarefa muito pesada, dada a vasta gama de situações envolvidas), mas apenas com o intuito de destacar alguns problemas, no âmbito da econometria, que o "Bootstrap" parece ultrapassar satisfatoriamente.

A) A estimação de distribuições por amostragem.

Antes de entrar, propriamente, no domínio econométrico, convirá realçar que uma das aplicações mais gerais do "Bootstrap" é a de estimar a distribuição por amostragem de uma estatística,

$$R \equiv R(X_1, X_2, ..., X_n ; F). \qquad (2.31)$$

Repare-se que $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n) \equiv \hat{\theta}(X_1, X_2, ..., X_n ; F)$ é caso particular da estatística R (a estatística R não tem, necessariamente, de ser um estimador de um parâmetro).

A aproximação "Bootstrap" de R é,

$$R^* \equiv R(X_1^*, X_2^*, ..., X_n^* ; \hat{F}). \tag{2.32}$$

A distribuição por amostragem de R^* pode ser aproximada pela distribuição empírica dos valores $\{R^{*b}, b = 1, 2, ..., B\}$, onde,

$$R^{*b} \equiv R(x_1^{*b}, x_2^{*b}, ..., x_n^{*b}), b = 1, 2, ..., B. \tag{2.33}$$

Então, como a distribuição empírica dos valores $\{R^{*b}, b = 1, 2, ..., B\}$ aproxima a distribuição por amostragem de R^* e R^* é a aproximação "Bootstrap" de R, pode dizer-se que a distribuição empírica dos valores $\{R^{*b}, b = 1, 2, ..., B\}$ é a estimativa da distribuição por amostragem de R.

B) A estimação de modelos de regressão.

No domínio da econometria uma das aplicações mais interessantes do "Bootstrap" reside na estimação de modelos de regressão múltipla. Vai apresentar-se o caso de modelos de uma só equação, embora também existam estudos de aplicação a modelos de equações simultâneas [veja-se Freedman (1984)].

Considere-se o seguinte modelo, de uma só equação, com n observações,

$$Y = g(X,\beta) + \varepsilon, \tag{2.34}$$

onde,

$$Y = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_n \end{bmatrix}^T ,$$

$$X = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \dots & \dots & \dots & \dots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_1 & \beta_2 & \dots & \beta_p \end{bmatrix}^T ,$$

$$g(X,\beta) = \begin{bmatrix} g_1(X_{1\bullet},\beta) & g_2(X_{2\bullet},\beta) & \dots & g_n(X_{n\bullet},\beta) \end{bmatrix}^T ,$$

$$\varepsilon = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_n \end{bmatrix}^T .$$

Repare-se que:

- Y é uma matriz coluna com as n observações da variável endógena.

  Y_i - i-ésima observação da variável endógena, i = 1, 2, ..., n.

- X é uma matriz com as n observações de cada uma das k variáveis exógenas.

  X_{ij} - i-ésima observação da j-ésima variável exógena, i = 1, 2, ..., n; j = 1, 2, ..., k.

- β é uma matriz coluna com p parâmetros.

- $g(X,\beta)$ é uma matriz coluna com as n observações da relação entre as variáveis exógenas e os parâmetros (a qual pode ser linear ou não).

  $g_i(X_{i\bullet},\beta)$ - i-ésima observação da relação entre as variáveis exógenas e os parâmetros, i = 1, 2, ..., n.

  Note-se que $X_{i\bullet}$ é a linha i da matriz X, i = 1, 2, ..., n.

- $\mathcal{E}$ é uma matriz coluna com os n valores da variável residual (variável aleatória não observável).

  ε_i - valor da variável residual correspondente à i-ésima equação, i = 1, 2, ..., n.

Sabe-se que,

$$\varepsilon_i \overset{iid}{\sim} F\ ,\ i = 1, 2, ..., n; \qquad E_F(\varepsilon_i) = 0, i = 1, 2, ..., n. \tag{2.35}$$

É comum assumir-se que a distribuição da variável residual (aqui representada pela sua função de distribuição F) segue uma lei normal de média nula e variância σ^2. Em muitos casos, esta hipótese não assenta em pressupostos correctos e destina-se, apenas, a permitir a boa continuação do estudo, a qual ficaria comprometida (nomeadamente, ficariam prejudicadas todas as inferências a realizar sobre β, a partir da distribuição do seu estimador) se, mais realisticamente, se assumisse o desconhecimento da distribuição da variável residual.

A aplicação do "Bootstrap" vem permitir a assumpção desse desconhecimento, sem problemas de maior - suponha-se, então, que F é desconhecida.

O objectivo é estimar os parâmetros β_i, i = 1, 2, ..., p, e sobre eles realizar determinadas inferências, o que se consegue com o seguinte estimador,

$$\hat{\beta} = \min_{\beta} D[Y, g(X,\beta)], \tag{2.36}$$

onde D é uma função que mede uma qualquer distância entre Y e $g(X,\beta)$ - a estimativa de β é o valor de β que minimiza a distância entre Y e $g(X,\beta)$.

O algoritmo "Bootstrap" para deduzir a distribuição do estimador $\hat{\beta}$, tem os seguintes passos:

1º - Começa-se por estimar β, de acordo com algum dos métodos habituais, supondo a normalidade das variáveis residuais.

Representando a estimativa de β por $\hat{\beta}$ [5], tem-se,

$$Y = g(X,\hat{\beta}) + \hat{\varepsilon} \qquad (2.37)$$

onde $\hat{\varepsilon}$ é uma matriz coluna com os valores estimados da variável residual, nas n equações, cada uma correspondente a uma observação.

2º - A partir da amostra dos ε_i - $(\hat{\varepsilon}_1, \hat{\varepsilon}_2, ..., \hat{\varepsilon}_n)$ - fazem-se tiragens com reposição, construindo B amostras "Bootstrap" (B suficientemente grande), de dimensão n,

$$(\hat{\varepsilon}^{*b}_1, \hat{\varepsilon}^{*b}_2, ..., \hat{\varepsilon}^{*b}_n),\ b = 1, 2, ..., B.$$

Para cada amostra "Bootstrap" $(\hat{\varepsilon}^{*b}_1, \hat{\varepsilon}^{*b}_2, ..., \hat{\varepsilon}^{*b}_n)$, b = 1, 2, ..., B, gera-se a correspondente amostra "Bootstrap" $(Y^{*b}_1, Y^{*b}_2, ..., Y^{*b}_n)$, fazendo,

$$Y^{*b} = g(X, \hat{\beta}) + \hat{\varepsilon}^{*b}, \qquad (2.38)$$

onde,

$$Y^{*b} = \left[Y^{*b}_1\ Y^{*b}_2\ ...\ Y^{*b}_n\right]^T \text{ e } \hat{\varepsilon}^{*b} = \left[\hat{\varepsilon}^{*b}_1\ \hat{\varepsilon}^{*b}_2\ ...\ \hat{\varepsilon}^{*b}_n\right]^T.$$

---

[5] Para não sobrecarregar a notação, vai representar-se, quer o estimador de β, quer a estimativa de β por $\hat{\beta}$. Ao longo do texto será indicado quando se trata do estimador ou da estimativa, tornando claro quando se está num ou noutro caso.

Repare-se que a extracção com reposição, a partir da amostra $(\hat{\varepsilon}_1, \hat{\varepsilon}_2, ..., \hat{\varepsilon}_n)$, equivale a substituir (2.35) por,

$$\hat{\varepsilon}_i^* \overset{iid.}{\sim} \hat{F}, \; i = 1, 2, ..., n, \tag{2.39}$$

onde os $\hat{\varepsilon}_i^*$ são aqui entendidos, não como estimativas concretas, mas sim como variáveis aleatórias "Bootstrap" que podem assumir qualquer valor de entre os valores da amostra observada $(\hat{\varepsilon}_1, \hat{\varepsilon}_2, ..., \hat{\varepsilon}_n)$.

3º - Para cada amostra "Bootstrap" $(Y_1^{*b}, Y_2^{*b}, ..., Y_n^{*b})$, $b = 1, 2, ..., B$, estima-se β da seguinte forma,

$$\hat{\beta}^{*b} = \min_\beta D[Y^{*b}, g(X,\beta)]. \tag{2.40}$$

A distribuição empírica dos valores "Bootstrap" $\{\hat{\beta}^{*1}, \hat{\beta}^{*2}, ..., \hat{\beta}^{*B}\}$, aproxima a distribuição por amostragem do estimador $\hat{\beta}$, permitindo realizar inferências estatísticas sobre os parâmetros β.

O inconveniente desta versão do "Bootstrap" no domínio da regressão, reside no facto de poder dar falsos bons resultados, em modelos sobreparametrizados [veja-se Efron (1982a) - pg. 36].

Outra versão ou forma de aplicar o "Bootstrap" à regressão, tem a ver com uma modificação do algoritmo: em vez de se efectuarem tiragens com reposição da amostra $(\hat{\varepsilon}_1, \hat{\varepsilon}_2, ..., \hat{\varepsilon}_n)$, opta-se por efectuar tiragens com reposição a partir dos valores observados da variável endógena e das variáveis exógenas. Sendo Y^{*b} e X^{*b} os valores "Bootstrap" de Y e X, respectivamente, os valores $\hat{\beta}^{*1}, \hat{\beta}^{*2}, ..., \hat{\beta}^{*B}$, obtêm-se, fazendo,

$$\hat{\beta}^{*b} = \min_\beta D[Y^{*b}, g(X^{*b},\beta)]. \tag{2.41}$$

Estas duas formas de aplicar o "Bootstrap" à regressão são assintoticamente equivalentes, mas podem dar resultados muito diferentes em pequenas amostras [veja-se Efron e Tibshirani (1986) - pg. 64].

C) A estimação do valor esperado do excesso de erro.

Outra importante aplicação do "Bootstrap", nos domínios da regressão, diz respeito à estimativa do valor esperado do excesso de erro.

Considerando o modelo (2.34), suponha-se que pretende prever-se o valor da variável endógena para o momento m, m > n (admite-se que o modelo é cronológico), ou seja, prever Y_m, com base em valores conhecidos (pré--fixados) para as variáveis exógenas, no momento m.

Seja,

$$X_m = [X_{m1}\ X_{m2}\ \dots\ X_{mk}],$$

com os valores pré-fixados para as k variáveis exógenas, no momento m.

A estimativa de Y_m é dada por,

$$\hat{Y}_m = g(X_m, \hat{\beta}), \qquad (2.42)$$

onde $\hat{\beta}$ representa a estimativa de β.

Ao estimar Y_m através de (2.42), comete-se um erro, porque o valor efectivamente assumido (ou melhor, a assumir) pela variável endógena é Y_m, enquanto que o valor previsto é $\hat{Y}_m$. Uma medida deste erro pode ser dada pela função,

$$Q(Y_m, \hat{Y}_m) = Q[Y_m, g(X_m, \hat{\beta})]. \qquad (2.43)$$

O excesso de erro não é mais do que a diferença entre duas parcelas:

- O valor esperado do verdadeiro erro cometido, $E_{mF}\{Q[Y_m, g(X_m, \hat{\beta})]\}$.

  Os índices m e F no valor esperado significam que este depende de F (relembre-se que $\varepsilon_i \overset{iid.}{\sim} F$, i = 1, 2, ..., n, e que a variável endógena depende da variável residual) e que está a ser calculado em relação ao período m.

  Tenha-se em atenção que, no cálculo deste valor esperado, $\hat{\beta}$ deve ser entendido como o estimador de β (e não como a sua estimativa) - a distribuição do estimador $\hat{\beta}$ depende de F, daí que o valor esperado dependa, também, de F.

- O valor esperado do erro aparente cometido, $E_{m\hat{F}}\{Q[Y_m, g(X_m, \hat{\beta})]\}$.

  Este valor esperado tem um significado semelhante ao anterior, com a diferença de se considerar $\hat{F}$ em vez de F - como a distribuição F é desconhecida, não se consegue calcular $E_{mF}\{Q[Y_m, g(X_m, \hat{\beta})]\}$, adoptando-se, como sua estimativa, $E_{m\hat{F}}\{Q[Y_m, g(X_m, \hat{\beta})]\}$, onde,

$$E_{m\hat{F}}\{Q[Y_m, g(X_m, \hat{\beta})]\} = \frac{1}{n} \sum_{j=1}^{n} Q[Y_j, g(X_{j\bullet}, \hat{\beta})]. \qquad (2.44)$$

  Tenha-se em atenção que, ao contrário do valor esperado $E_{mF}\{Q[Y_m, g(X_m, \hat{\beta})]\}$, onde $\hat{\beta}$ representa o estimador de β, no valor esperado $E_{m\hat{F}}\{Q[Y_m, g(X_m, \hat{\beta})]\}$, $\hat{\beta}$ representa a estimativa de β.

  A justificação para (2.44), assenta no facto de a distribuição empírica da amostra (cuja função de distribuição é $\hat{F}$) atribuir probabilidade

1/n a cada "ponto amostra" formado pela j-ésima observação das variáveis exógenas e da variável endógena: $(X_{j1}, X_{j2}, ..., X_{jk}, Y_j)$, j = 1, 2, ..., n.

Então, o excesso de erro pode ser representado por,

$$T \equiv T(X,F) = E_{mF}\{Q[Y_m, g(X_m, \hat{\beta})]\} - E_{m\hat{F}}\{Q[Y_m, g(X_m, \hat{\beta})]\}. \qquad (2.45)$$

O objectivo é estimar o valor esperado do excesso de erro, ou seja, estimar,

$$EEE \equiv E_F[T(X,F)], \qquad (2.46)$$

onde o índice F significa que este valor esperado depende de F.

O "Bootstrap" resolve este problema fazendo a aproximação "Bootstrap" de T,

$$T^* \equiv T(X^*,\hat{F}) = E_{m\hat{F}}\{Q[Y_m, g(X_m, \hat{\beta}^*)]\} - E_{m\hat{F}^*}\{Q[Y_m, g(X_m, \hat{\beta}^*)]\}, \qquad (2.47)$$

onde,

- $\hat{\beta}^*$ é a estimativa "Bootstrap" de β, obtida com X^* e Y^*,

$$\hat{\beta}^* = \min_{\beta} D[Y^*, g(X^*, \beta)]. \qquad (2.48)$$

Por sua vez, X^* é a matriz X "Bootstrap", a qual se obtem fazendo tiragens com reposição a partir dos valores observados das variáveis exógenas (para os momentos 1, 2, ..., n) e Y^* é a matriz Y "Bootstrap", a qual se obtem fazendo tiragens com reposição a partir dos valores observados da variável endógena (para os momentos 1, 2, ..., n).

- $\hat{F}^*$ é a função de distribuição empírica "Bootstrap" da amostra observada (X,Y). Note-se que a distribuição empírica "Bootstrap" da amostra atribui probabilidade,

$$P_j^* = \frac{\#\{(X_{i1}^*, X_{i2}^*, ..., X_{ik}^*, Y_i^*) = (X_{j1}, X_{j2}, ..., X_{jk}, Y_j)\}}{n},$$

a cada "ponto amostra" $(X_{j1}, X_{j2}, ..., X_{jk}, Y_j)$, $j = 1, 2, ..., n$.

Atendendo a que $E_{m\hat{F}}\{Q[Y_m, g(X_m, \hat{\beta}^*)]\} = \sum_{j=1}^{n} \frac{1}{n} Q[Y_j^*, g(X_{j\bullet}^*, \hat{\beta}^*)]$, onde $X_{j\bullet}^* = (X_{j1}^*, X_{j2}^*, ..., X_{jk}^*)$, e $E_{m\hat{F}^*}\{Q[Y_m, g(X_m, \hat{\beta}^*)]\} = \sum_{j=1}^{n} P_j^* \; Q[Y_j^*, g(X_{j\bullet}^*, \hat{\beta}^*)]$, pode reescrever-se (2.47),

$$T^* \equiv T(X^*, \hat{F}) = \sum_{j=1}^{n} \left(\frac{1}{n} - P_j^*\right) Q[Y_j^*, g(X_{j\bullet}^*, \hat{\beta}^*)]. \qquad (2.49)$$

Fazendo $P_j^\circ = 1/n$, o valor esperado de (2.49), $E_*(T^*)$ (onde o índice $*$ indica que o valor esperado é calculado com base na distribuição empírica da amostra "Bootstrap", representada por $\hat{F}^*$), é dado por,

$$E_*(T^*) = E_*\left\{\sum_{j=1}^{n} (P_j^\circ - P_j^*) \; Q[Y_j^*, g(X_{j\bullet}^*, \hat{\beta}^*)]\right\}. \qquad (2.50)$$

Como T^* é a aproximação "Bootstrap" de T, $E_*(T^*)$ é o estimador "Bootstrap" de $EEE \equiv E_F[T(X,F)]$,

$$\hat{EEE}_{BOOT} = E_*(T^*). \qquad (2.51)$$

D) O caso das subamostras com distribuição diferente.

Para terminar estas breves ilustrações sobre a aplicabilidade do "Bootstrap", foca-se um problema com que se depara em alguns estudos econométricos: o universo que estamos a analisar não se caracteriza por uma só distribuição aleatória, pelo contrário, o espaço de resultados ou espaço amostra, χ, encontra-se particionado em H subespaços amostra, χ_h,

$$\chi = \sum_{h=1}^{H} \chi_h . \tag{2.52}$$

A distribuição a que obedecem os elementos do universo difere, consoante o subespaço amostra a que pertencem,

$$X_{ih} \overset{iid.}{\sim} F_h , X_{ih} \in \chi_h , h = 1, 2, ..., H, \tag{2.53}$$

onde,

F_h - função de distribuição (desconhecida) de cada X_{ih}, $X_{ih} \in \chi_h$.

O objectivo é realizar inferências sobre um parâmetro unidimensional θ, o qual depende das funções de distribuição F_h, $h = 1, 2, ..., H$,

$$\theta \equiv \theta(F_1, F_2, ..., F_H). \tag{2.54}$$

Construa-se uma amostra aleatória de dimensão n,

$$(X_{ih}, i = 1, 2, ..., n_h, h = 1, 2, ..., H), \tag{2.55}$$

onde $\sum_{h=1}^{H} n_h = n$.

Tendo-se observado $X_{ih} = x_{ih}$, $i = 1, 2, ..., n_h$, $h = 1, 2, ..., H$, podem-se construir as H funções de distribuição empíricas, as quais são dadas por,

$$\hat{F}_h \equiv \hat{F}_h(x) = \frac{\#\{x_{ih} \leq x\}}{n_h}, \quad -\infty < x < +\infty, \; h = 1, 2, ..., H. \tag{2.56}$$

Para fazer inferência sobre o parâmetro $\theta \equiv \theta(F_1, F_2, ..., F_H)$, utiliza-se um estimador,

$$\hat{\theta} \equiv \hat{\theta}(X_{ih}, i = 1, 2, ..., n_h, h = 1, 2, ..., H), \tag{2.57}$$

cuja distribuição depende das distribuições do universo F_h, $h = 1, 2, ..., H$. O desconhecimento destas, leva o "Bootstrap" a substituir F_h por $\hat{F}_h$, $h = 1, 2, ..., H$ - na prática (e dado tratar-se de um caso não paramétrico), isto equivale a realizar B tiragens com reposição (B suficientemente grande), a partir de cada uma das H particulares subamostras observadas, (x_{ih}, $i = 1, 2, ..., n_h$, $h = 1, 2, ..., H$), gerando-se B amostras "Bootstrap". Na posse destas amostras, podem calcular-se os valores "Bootstrap" do estimador $\hat{\theta}$ - $\hat{\theta}^{*1}$, $\hat{\theta}^{*2}$, ..., $\hat{\theta}^{*B}$ - desenrolando-se tudo o resto tal como foi exposto no ponto 2.1 deste trabalho.

Como já foi dito, as aplicações do "Bootstrap" aqui expostas representam apenas uma pequena parte dos domínios em que este método revela grandes potencialidades. Para outras aplicações, podem-se consultar vários artigos como, por exemplo, nos domínios da econometria: Flood (1985), sobre os modelos Tobit; Freedman (1984), sobre modelos lineares de equações simultâneas; Härdle e Bowman (1988), sobre a regressão não linear; Wu (1986), sobre vários problemas ligados à regressão; entre outros.

Julga-se que a importância do "Bootstrap" e os seus largos horizontes de aplicação ficaram bem expressos, através da breve abordagem aqui efectuada.

2.4 - A VALIDADE ASSINTÓTICA DO "BOOTSTRAP"

A teoria assintótica do "Bootstrap" é um "mundo" complexo e difícil de penetrar e constitui, por si só, um tema susceptível de se tratar noutro trabalho. No presente, enumeram-se somente alguns resultados que mostram a validade assintótica do "Bootstrap".

Assuma-se uma situação definida por (2.1), (2.2), (2.3) e (2.4) e considere-se a variável aleatória,

$$H \equiv \sqrt{n}\left[\hat{\theta}(X_1, X_2, ..., X_n) - \theta(F)\right] \equiv \sqrt{n}\,(\hat{\theta} - \theta). \qquad (2.58)$$

Supondo que o estimador $\hat{\theta}(X_1, X_2, ..., X_n)$ permite a aplicação do Teorema do Limite Central, tem-se,

$$H \equiv \sqrt{n}\left[\hat{\theta}(X_1, X_2, ..., X_n) - \theta(F)\right] \dot{\sim} N(0,\omega^2), \qquad (2.59)$$

onde,

- O símbolo $\dot{\sim}$ significa "tem uma distribuição assintótica".

- $N(0,\omega^2)$ é a distibuição normal, com média nula e variância ω^2.

Designe-se por $H_n(\cdot,F)$ a função de distribuição exacta da variável aleatória H.

O objectivo é estimar $H_n(\cdot,F)$, para assim poder realizar inferências sobre o parâmetro θ.

A estimativa de $H_n(\cdot,F)$, proposta pelo método "Bootstrap" é, naturalmente, $H_n(\cdot,\hat{F})$ - seguiu-se a filosofia "Bootstrap", ao substituir a distribuição do universo F, pela distribuição empírica da amostra $\hat{F}$.

O comportamento assintótico (tenha-se em atenção que o assintótico é em relação à dimensão da amostra casual, n) da estimativa "Bootstrap", $H_n(\cdot,\hat{F})$ foi estudado por diversos autores, podendo afirmar-se que $H_n(\cdot,\hat{F})$ é consistente no seguinte sentido: $H_n(\cdot,\hat{F})$ converge fracamente, em probabilidade, para a mesma distribuição normal para a qual converge $H_n(\cdot,F)$ (não se esqueça que $H_n(\cdot,F)$ é a função de distribuição da estatística H e que esta tem uma distribuição assintoticamente normal). Esta consistência prova-se, sob algumas hipóteses, aliás bastante gerais, como se pode ver em [Efron (1979) - pg. 23] e [Bickel e Freedman (1981) - pg. 1196-1199].

Outra propriedade assintótica importante, e particularmente interessante, da estimativa "Bootstrap", tem a ver com o facto desta ser, sob certas condições, assintoticamente minimax. Desenvolva-se um pouco esta ideia, com base na teoria da decisão estatística.

O problema consiste na estimação de $H_n(\cdot,F)$. Para o resolver, pode definir-se um conjunto de acções possíveis, onde cada acção é uma função de distribuição concreta proposta como estimativa da função de distribuição $H_n(\cdot,F)$,

$$A = \{H_n(\cdot,G): H_n(\cdot,G) \text{ é uma função de distribuição concreta proposta como estimativa da função de distribuição } H_n(\cdot,F)\}, \qquad (2.60)$$

e um espaço de estados, onde cada estado é uma função de distribuição possível de ser igual a $H_n(\cdot,F)$,

Θ = {$H_n(\cdot,G)$: $H_n(\cdot,G)$ é uma função de distribuição possível de ser igual a $H_n(\cdot,F)$}. (2.61)

Note-se que A = Θ , como é usual em problemas de estimação.

Como funções de decisão, têm-se as inúmeras estimativas $H_n(\cdot,F)$, entre as quais, a estimativa "Bootstrap",

$$\delta_{BOOT} = H_n(\cdot,\hat{F}). \tag{2.62}$$

Ao empregar uma função de decisão $\delta \equiv \delta(x_1, x_2, ..., x_n) = H_n(\cdot,G)$ (esta função de decisão equivale a propôr G como estimativa de F), incorre-se numa perca, que pode definir-se da seguinte forma,

$$L[H_n(\cdot,F), \delta] = L[H_n(\cdot,F), H_n(\cdot,G)] = u(\sqrt{n} \, \| H_{n,v}(\cdot,G) - H_{n,v}(\cdot,F) \|), \tag{2.63}$$

onde,

- u é uma função de $\mathbb{R}^+$ em $\mathbb{R}^+$, monótona crescente e limitada.

- O índice v em $H_{n,v}(\cdot,G)$ e $H_{n,v}(\cdot,F)$, significa convolução com a função de densidade de probabilidade $v(x) = (1 - |x|/a)\, a^{-1}$, se $|x| \leq a$. Mais propriamente, $H_{n,v}(\cdot,G)$ significa convolução de $H_n(\cdot,G)$ com v(x) e $H_{n,v}(\cdot,F)$ significa convolução de $H_n(\cdot,F)$ com v(x).

- A norma $\|\cdot\|$ define-se como, $\| h(x)\| = \sup_x |h(x)|$.

Esta função perca, pode interpretar-se como uma medida da distância entre a "verdadeira" função de distribuição $H_n(\cdot,F)$ e a sua estimativa proposta $H_n(\cdot,G)$, se se esquecer a convolução com v(x).

Mais pormenores sobre a construção de (2.63) podem ser vistos em [Beran (1982) - pg. 214-216].

Defina-se a função risco ou perca esperada,

$$R[H_n(\bullet,F), \delta] = R[H_n(\bullet,F), H_n(\bullet,G)] = E_F\{L[H_n(\bullet,F), H_n(\bullet,G)]\} =$$

$$= E_F[u(\sqrt{n} \parallel H_{n,v}(\bullet,G) - H_{n,v}(\bullet,F) \parallel)], \qquad (2.64)$$

a qual exprime a perca média sofrida pelo decisor quando emprega a função de decisão $\delta = H_n(\bullet,G)$ e o estado da natureza é $H_n(\bullet,F)$ [veja-se Murteira (1988a) - pg. 105].

Sob certas condições (as quais podem ser vistas em [Beran (1982) - pg. 215]), prova-se que a estimativa "Bootstrap", $H_n(\bullet,\hat{F})$ é assintoticamente minimax, ou seja, verifica,

$$\sup_F R[H_n(\bullet,F), H_n(\bullet,\hat{F})] = \inf_G \sup_F R[H_n(\bullet,F), H_n(\bullet,G)], \qquad (2.65)$$

quando $n \to +\infty$.

A justificação de (2.65) pode ver-se em [Beran (1982) - pg. 214-218]. Repare-se que não é dito que $H_n(\bullet,\hat{F})$ é a única estimativa de $H_n(\bullet,F)$ assintoticamente minimax - podem existir outras estimativas de $H_n(\bullet,F)$, também elas assintoticamente minimax [veja-se Beran (1982) - pg. 218].

O certo é que a estimativa de $H_n(\bullet,F)$, decorrente da aproximação à Normal expressa em (2.59) - $\Phi[\bullet/s_n(\hat{F})]$, onde Φ é a função de distribuição de uma normal estandardizada e $s_n(\hat{F})$ é uma estimativa de ω - não é, em geral, assintoticamente minimax [veja-se Beran (1982) - pg. 218], o que leva a concluir pela superioridade assintótica da aproximação "Bootstrap", no que diz respeito ao critério minimax, pelo menos na generalidade dos casos.

3 - O "BOOTSTRAP" NA CONSTRUÇÃO DE INTERVALOS DE CONFIANÇA

3.1 - "CASO PARAMÉTRICO" VERSUS "CASO NÃO PARAMÉTRICO"

No estudo da construção de intervalos de confiança "Bootstrap" vai dar-se especial ênfase ao caso não paramétrico, aquele em que a valia e a utilidade do "Bootstrap" se fazem sentir de uma forma mais acentuada.

Apesar de o estudo incidir especialmente sobre o caso não paramétrico, há que fazer referências ao caso paramétrico, já que, alguns dos resultados apresentados para o primeiro não são mais do que extrapolações de resultados deduzidos para o segundo. Estão neste caso a constante "a", conhecida por constante de aceleração (a qual desempenha um papel fundamental na construção de intervalos de confiança "Bootstrap" corrigidos do enviesamento e da aceleração da variância) e as propriedades dos intervalos de confiança "Bootstrap" com correcção do enviesamento e da aceleração da variância.

É natural, pelo menos de um ponto de vista intuitivo, que alguns resultados sobre intervalos de confiança "Bootstrap" não paramétricos sejam meras extrapolações ou conjecturas, efectuadas a partir de resultados deduzidos para intervalos de confiança "Bootstrap" paramétricos: a quase ausência de informação no caso não paramétrico traduz-se num campo de trabalho hostil, onde se torna muito difícil apresentar resultados mais ousados ou concludentes, daí que se recorra ao caso paramétrico, onde há uma maior abundância de informação, permitindo deduzir importantes resultados, os quais são, depois, extrapolados e adaptados (com as devidas reservas) ao caso não paramétrico. Esta a razão porque um trabalho sobre intervalos de

confiança "Bootstrap", em domínios não paramétricos, tem de conter referências, quase que obrigatórias, ao caso paramétrico.

3.2 - O MÉTODO DOS PERCENTIS

Considere-se a situação definida por (2.1), (2.2), (2.3) e (2.4), em que a função de distribuição, F, é desconhecida (está-se no caso não paramétrico). O propósito é construir um intervalo de confiança a $(1-2\alpha)100\%$, $0 < \alpha < 0.5$, para o parâmetro $\theta \equiv \theta(F)$. Por outras palavras, pretende-se determinar, θ_{LI} e θ_{LS}, tais que o intervalo,

$$[\theta_{LI} ; \theta_{LS}], \tag{3.1}$$

contenha o verdadeiro valor do parâmetro, $\theta \equiv \theta(F)$, com $(1-2\alpha)100\%$ de confiança.

Sendo a distribuição do universo desconhecida, os limites inferior e superior de (3.1), θ_{LI} e θ_{LS}, respectivamente, nunca serão exactamente conhecidos, tendo de ser estimados. Assim, o problema central deste trabalho vai girar em torno dos estimadores para θ_{LI} e θ_{LS}, apresentando-se algumas soluções, baseadas em diferentes hipóteses, mas todas elas relacionadas com o "Bootstrap".

Antes de apresentar o método dos percentis (a primeira aplicação do "Bootstrap" aos intervalos de confiança), recorda-se a tradicional solução para construir o intervalo de confiança (3.1).

Suponha-se que o estimador $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$ é o estimador da máxima verosimilhança de θ. Prova-se que, se $f(x) \equiv f(x;\theta)$ (está-se a considerar o caso paramétrico, em que a função de densidade de probabilidade

- ou função de probabilidade - da distribuição do universo depende de um só parâmetro θ) satisfizer certas condições de regularidade [as quais podem ser vistas em Murteira (1980) - pg. 165], tem-se,

$$\hat{\theta} \overset{\cdot}{\sim} N[\theta, 1/I(X_1, X_2, ..., X_n ; \theta)], \tag{3.2}$$

onde $I(X_1, X_2, ..., X_n ; \theta) = n\, E\left\{\left[\frac{\partial}{\partial\theta} \ln f(X;\theta)\right]^2\right\}$ é a quantidade de informação de Fisher [sobre (3.2) veja-se Murteira (1980) - pg. 166].

A partir de (3.2) e fazendo,

$$\hat{\sigma}_{IF} = \sqrt{\frac{1}{I(X_1, X_2, ..., X_n ; \theta)}}, \tag{3.3}$$

vem,

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}_{IF}} \overset{\cdot}{\sim} N(0,1). \tag{3.4}$$

Com base em (3.4), pode dizer-se que,

$$\Phi(z_\alpha) = \alpha,\ \Phi(z_{1-\alpha}) = 1-\alpha \Leftrightarrow P\left(z_\alpha < \frac{\hat{\theta} - \theta}{\hat{\sigma}_{IF}} < z_{1-\alpha}\right) \cong 1-2\alpha, \tag{3.5}$$

De (3.5) vem,

$$P\left(\hat{\theta} - \hat{\sigma}_{IF}\, z_{1-\alpha} < \theta < \hat{\theta} - \hat{\sigma}_{IF}\, z_\alpha\right) \cong 1-2\alpha, \tag{3.6}$$

donde se conclui que os estimadores tradicionais para θ_{LI} e θ_{LS} são, respectivamente,

$$\hat{\theta}_{TRA}(LI) = \hat{\theta} - \hat{\sigma}_{IF}\, z_{1-\alpha}, \tag{3.7}$$

e,

$$\hat{\theta}_{TRA}(LS) = \hat{\theta} - \hat{\sigma}_{IF}\, z_{\alpha} . \qquad (3.8)$$

Assim sendo, um intervalo de confiança a $(1-2\alpha)100\%$ (aproximadamente) para θ, será dado por,

$$\left[\hat{\theta}_{TRA}(LI) \; ; \; \hat{\theta}_{TRA}(LS)\right]. \qquad (3.9)$$

Este método tradicional sofre de alguns inconvenientes:

- É um método que se baseia em resultados assintóticos, logo só tem validade em amostras suficientemente grandes.

- Mesmo que se obtenham bons resultados, em grandes amostras, na construção de intervalos de confiança para θ, o mesmo já não se pode afirmar em relação a intervalos de confiança para determinadas funções de θ. Por outras palavras, mesmo que (3.9) seja um "bom" intervalo de confiança para θ, isso não quer dizer que $\left[h[\hat{\theta}_{TRA}(LI)] \; ; \; h[\hat{\theta}_{TRA}(LS)]\right]$ seja um bom intervalo de confiança para a função de θ, $h(\theta)$ [veja-se Efron (1987) - pg. 171].

- Finalmente, é um método que se aplica em domínios paramétricos, mas não a casos não paramétricos de grande ausência de informação. Nesta última situação, o resultado (3.2) não adianta muito, devido à dificuldade em se calcular (3.3). O mesmo se pode dizer, se se recorrer a um resultado mais geral do que (3.2) - o Teorema de Lindeberg--Levy [veja-se Murteira (1979) - pg. 354], pelo qual,

$$\frac{\hat{\theta} - \theta(F)}{\sigma(F)} \overset{\cdot}{\sim} N(0,1), \qquad (3.10)$$

onde $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, \ldots, X_n)$ é um estimador de $\theta \equiv \theta(F)$, tendo em atenção que este resultado continuaria a ser válido (em termos

assintóticos), substituindo $\sigma(F)$ por um estimador consistente. Mantem-se o problema em calcular o estimador consistente de $\sigma(F)$, para além de o resultado (3.10) não ser válido para todas as variáveis aleatórias $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$, mas apenas para aquelas que se possam exprimir como uma soma de variáveis aleatórias independentes e idênticamente distribuídas, com média e variância finitas.

Como se vê, a abordagem tradicional enferma de algumas limitações, para além de se tornar muito difícil (em certos casos até impossível) a sua aplicação aos domínios não paramétricos. Veja-se, então, como o "Bootstrap" permite ultrapassar este impasse.

Desejando-se um intervalo de confiança a $(1-2\alpha)100\%$ para $\theta \equiv \theta(F)$, a ideia consiste em utilizar a função de distribuição do estimador $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$ para construir esse mesmo intervalo. No entanto, o problema reside em que a função de distribuição de $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$ é desconhecida. O método dos percentis, inspirado no "Bootstrap", permite solucionar este problema, construindo a função de distribuição "Bootstrap" de $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n) \equiv \hat{\theta}(X_1, X_2, ..., X_n ; F)$, a qual não é mais do que a função de distribuição de,

$$\hat{\theta}^* \equiv \hat{\theta}(X_1^*, X_2^*, ..., X_n^* ; \hat{F}). \qquad (3.11)$$

A função de distribuição de (3.11) obtem-se, de forma aproximada, com base no seguinte algoritmo de Monte Carlo:

1º) A partir da particular amostra observada $(x_1, x_2, ..., x_n)$, realiza-se uma extracção com reposição, por forma a construir uma nova amostra de dimensão n, $(x_1^*, x_2^*, ..., x_n^*)$.

2º) Calcula-se o valor "Bootstrap" de $\hat{\theta}$, $\hat{\theta}^* \equiv \hat{\theta}(x_1^*, x_2^*, ..., x_n^*)$.

3º) Repetem-se os passos 1º e 2º B vezes, onde B é suficientemente grande.

Assim, fica-se com um conjunto de B valores "Bootstrap" de $\hat{\theta}$ - $\hat{\theta}^{*1}$, $\hat{\theta}^{*2}$, ..., $\hat{\theta}^{*B}$, onde $\hat{\theta}^{*b} \equiv \hat{\theta}(x_1^{*b}, x_2^{*b}, ..., x_n^{*b})$, b = 1, 2, ..., B.

4º) A função de distribuição empírica "Bootstrap" de $\hat{\theta}$ constrói-se facilmente,

$$\hat{G}(t) = \frac{\#\{\hat{\theta}^{*b} \leq t\}}{B}, \quad 1 \leq b \leq B, \quad -\infty < t < +\infty. \tag{3.12}$$

Note-se que a seguinte igualdade,

$$\hat{G}(t) = P_{\hat{F}}(\hat{\theta}^* \leq t) \equiv P_*(\hat{\theta}^* \leq t) \equiv P_*[\hat{\theta}(X_1^*, X_2^*, ..., X_n^* ; \hat{F}) \leq t]$$

só é válida quando $B \to +\infty$, ou seja, $\hat{G}(t)$ não é a verdadeira função de distribuição "Bootstrap" de $\hat{\theta}$, já que, não se conhece a exacta distribuição de (3.11) - a verdadeira distribuição "Bootstrap" de $\hat{\theta}$ - mas sim apenas uma sua estimativa.

Os estimadores de θ_{LI} e θ_{LS}, dados pelo método dos percentis, são, respectivamente,

$$\hat{\theta}_{PER}(LI) = \hat{G}^{-1}(\alpha) = t_0 : t_0 = \inf_t \frac{\#\{\hat{\theta}^{*b} \leq t\}}{B} \geq \alpha, \tag{3.13}$$

e,

$$\hat{\theta}_{PER}(LS) = \hat{G}^{-1}(1-\alpha) = t_1 : t_1 = \inf_t \frac{\#\{\hat{\theta}^{*b} \leq t\}}{B} \geq 1-\alpha. \tag{3.14}$$

Assim, o intervalo de confiança a $(1-2\alpha)100\%$ (aproximadamente) para θ, dado pelo método dos percentis, é,

$$\left[\hat{\theta}_{PER}(LI)\ ;\ \hat{\theta}_{PER}(LS)\right] = \left[\hat{G}^{-1}(\alpha)\ ;\ \hat{G}^{-1}(1-\alpha)\right]. \qquad (3.15)$$

As diferenças $\left|\hat{\theta}_{PER}(LI) - \theta_{LI}\right|$ e $\left|\hat{\theta}_{PER}(LS) - \theta_{LS}\right|$ são $O_P(n^{-1})$ (pelo menos, em domínios paramétricos unidimensionais, podendo conjecturar-se que também o sejam em domínios não paramétricos) [veja-se Dicicio e Romano (1988) - pg. 340], pelo que, para pequenas amostras, (3.15) pode ser significativamente diferente de (3.1) que é o intervalo de confiança exacto, a $(1-2\alpha)100\%$, para θ. Este facto é o reflexo de algumas limitações do método dos percentis, as quais motivam a busca de outros métodos "Bootstrap" para construir o desejado intervalo de confiança. Neste sentido, surgem o método dos percentis corrigido do enviesamento e o método dos percentis corrigido do enviesamento e da aceleração da variância, os quais procuram dar resposta às limitações do método dos percentis, com base na teoria da transformação. Esta matéria vai analisar-se nas páginas seguintes.

3.3 - A MELHORIA DO MÉTODO DOS PERCENTIS COM BASE NA TEORIA DA TRANSFORMAÇÃO

3.3.1 - TÓPICOS SOBRE A TEORIA DA TRANSFORMAÇÃO

Como se disse, no ponto anterior do trabalho, o método dos percentis pode ser significativamente melhorado, com recurso à teoria da transformação. A presente secção pretende ser uma pequena introdução sobre a teoria da transformação e focar os aspectos que mais interessam para o tratamento dos intervalos de confiança "Bootstrap".

Os resultados a apresentar vão ser deduzidos com base numa família de distribuições contínuas, com parâmetro unidimensional. Como se está a tratar os intervalos de confiança "Bootstrap" em domínios não paramétricos, os resultados aqui apresentados, para o caso paramétrico, serão alargados para o caso não paramétrico.

Considere-se uma família de distribuições contínuas paramétricas, com função de distribuição representada por $M \equiv M_\gamma(x)$,

$$\mathfrak{M} = \{M_\gamma(x), \gamma \in \Gamma\}, \tag{3.16}$$

onde γ é um parâmetro unidimensional e Γ o espaço dos parâmetros.

Seja X uma variável aleatória contínua, cuja "verdadeira" distribuição se encontra entre as distribuições da família (3.16): existirá alguma função monótona, g, que normalize a família $\mathfrak{M}$? Formalizando, trata-se de saber se existe uma transformação monótona, g, tal que, sendo Y = g(X), a variável aleatória Y tenha uma distribuição aproximadamente normal, isto para todas as variáveis aleatórias X que pertençam à família $\mathfrak{M}$.

Para resolver o problema, defina-se,

$$D(z,\gamma) = \frac{m_\gamma[x_{\Phi(z),\gamma}]}{m_\gamma[x_{0.5,\gamma}]} \frac{\phi(0)}{\phi(z)}, \qquad (3.17)$$

como sendo uma função diagnóstico [veja-se Efron (1982b) - pg. 325], onde,

$$m_\gamma(x) \equiv m(x|\gamma) \equiv \frac{\partial}{\partial\gamma} M_\gamma(x).$$

$x_{\alpha,\gamma} \equiv x : M_\gamma(x) = \alpha$, $0 < \alpha < 1$ - α-ésimo quantil da variável aleatória X.

$\phi(z)$ - função de densidade de probabilidade normal estandardizada.

A motivação para esta função diagnóstico assenta na seguinte transformação local para a normalidade,

$$t_\gamma(X) = \Phi^{-1}[M_\gamma(X)] \sim N(0,1), \qquad (3.18)$$

resultado que permite reescrever (3.17),

$$D(z,\gamma) = \frac{\frac{\partial}{\partial\gamma} t_\gamma[x_{\Phi(z),\gamma}]}{\frac{\partial}{\partial\gamma} t_\gamma[x_{0.5,\gamma}]}. \qquad (3.19)$$

A função diagnóstico irá desempenhar um importante papel na normalização da família $\mathfrak{M}$, pois prova-se que assume determinada forma quando a família $\mathfrak{M}$ for normalizável. Antes de entrar neste aspecto, defina-se o conceito de "família de transformações gerais escaladas", ou abreviadamente GSTF (iniciais do inglês "general scaled transformation family").

Diz-se que $\mathfrak{M}$ é uma GSTF, se existir uma função estritamente monótona, g, tal que,

$$\frac{g(X) - \nu_\gamma}{\sigma_\gamma} \sim q(Z), \tag{3.20}$$

onde,

- $Z \sim N(0,1)$;

- $q(Z)$ é uma função estritamente crescente e diferenciável, verificando $q(0) = 0$ e $q'(0) = 1$;

- ν_γ e $\sigma_\gamma > 0$ são funções de γ diferenciáveis, mas não necessariamente monótonas;

- $\frac{\partial \nu_\gamma}{\partial \gamma} \neq 0$, excepto para um número finito de valores de γ;

[sobre este resultado e o lema seguinte veja-se Efron (1982b) - pg. 326].

Os resultados (3.17) e (3.20) permitem enunciar o seguinte:

Lema 1 - Se $\mathfrak{M}$ for uma GSTF, então,

$$D(z,\gamma) = \frac{1 + q(z)\ \varepsilon_\gamma}{q'(z)}, \text{ onde } \varepsilon_\gamma = \frac{\partial \sigma_\gamma}{\partial \gamma} \Big/ \frac{\partial \nu_\gamma}{\partial \gamma}.$$

Para efectuar a demonstração, faça-se $\tilde{Z} = q(Z)$. A função de distribuição de $\tilde{Z}$ é dada por,

$$\tilde{\Phi}(\tilde{z}) = \Phi[q^{-1}(\tilde{z})], \tag{3.21}$$

porque $\tilde{z} = q(z) \Leftrightarrow z = q^{-1}(\tilde{z})$ [6].

Representem-se, por z_α , o quantil α da distribuição Z, e, por $\tilde{z}_\alpha$, o quantil α da distribuição $\tilde{Z}$, ou seja,

$$z_\alpha : \Phi(z_\alpha) = \alpha, \tag{3.22}$$

$$\tilde{z}_\alpha : \tilde{\Phi}(\tilde{z}_\alpha) = \tilde{\Phi}[q(z_\alpha)] = \alpha. \tag{3.23}$$

A função de densidade de probabilidade de $\tilde{Z}$ representa-se por $\tilde{\phi}(\tilde{z}) = \frac{\partial}{\partial \tilde{z}} \tilde{\Phi}(\tilde{z})$ e satisfaz,

$$\tilde{\phi}(\tilde{z}) = \frac{\phi(z)}{q'(z)}, \tag{3.24}$$

$$\tilde{\phi}(0) = \frac{\phi(0)}{q'(0)} = \phi(0) = \frac{1}{\sqrt{2\pi}}. \tag{3.25}$$

Atendendo à função diagnóstico dada por (3.17), vai calcular-se $m_\gamma(x) \equiv \frac{\partial}{\partial \gamma} M_\gamma(x)$.

Tem-se,

$$M_\gamma(x) = P_\gamma(X \leq x) =$$

$$= P_\gamma[g(X) \leq g(x)] = P_\gamma\left[\frac{g(X) - \nu_\gamma}{\sigma_\gamma} \leq \frac{g(x) - \nu_\gamma}{\sigma_\gamma}\right],$$

---

[6] Note-se que z é um valor genérico assumido pela variável aleatória Z e $\tilde{z}$ é um valor genérico assumido pela variável aleatória $\tilde{Z}$.

isto é,

$$M_\gamma(x) = P_\gamma\left[\tilde{Z} \le \frac{g(x) - \nu_\gamma}{\sigma_\gamma}\right] = \tilde{\Phi}\left[\frac{g(x) - \nu_\gamma}{\sigma_\gamma}\right] \qquad (3.26)$$

Derivando (3.26) em ordem a γ, através da regra de derivação da função composta, sai,

$$m_\gamma(x) \equiv \frac{\partial}{\partial\gamma} M_\gamma(x) = \tilde{\phi}\left[\frac{g(x) - \nu_\gamma}{\sigma_\gamma}\right] \frac{-\dot{\nu}_\gamma\,\sigma_\gamma - \dot{\sigma}_\gamma\,[g(x) - \nu_\gamma]}{\sigma_\gamma^2} =$$

$$= \tilde{\phi}\left[\frac{g(x) - \nu_\gamma}{\sigma_\gamma}\right]\left[-\frac{\dot{\nu}_\gamma}{\sigma_\gamma} - \frac{g(x) - \nu_\gamma}{\sigma_\gamma}\frac{\dot{\sigma}_\gamma}{\sigma_\gamma}\right] =$$

$$= -\tilde{\phi}\left[\frac{g(x) - \nu_\gamma}{\sigma_\gamma}\right]\left[\frac{\dot{\nu}_\gamma}{\sigma_\gamma} + \frac{g(x) - \nu_\gamma}{\sigma_\gamma}\frac{\dot{\sigma}_\gamma}{\sigma_\gamma}\right], \qquad (3.27)$$

onde $\dot{\nu}_\gamma = \frac{\partial \nu_\gamma}{\partial\gamma}$ e $\dot{\sigma}_\gamma = \frac{\partial \sigma_\gamma}{\partial\gamma}$.

Fazendo $x = x_{\alpha,\gamma}$ e $\frac{g(x) - \nu_\gamma}{\sigma_\gamma} = \tilde{z}_\alpha$, em (3.27), chega-se a,

$$m_\gamma(x_{\alpha,\gamma}) \equiv \frac{\partial}{\partial\gamma} M_\gamma(x_{\alpha,\gamma}) = -\tilde{\phi}(\tilde{z}_\alpha)\left[\frac{\dot{\nu}_\gamma}{\sigma_\gamma} + \tilde{z}_\alpha\frac{\dot{\sigma}_\gamma}{\sigma_\gamma}\right].$$

Por (3.24), a última expressão pode escrever-se da seguinte forma,

$$m_\gamma(x_{\alpha,\gamma}) \equiv \frac{\partial}{\partial\gamma} M_\gamma(x_{\alpha,\gamma}) = -\frac{\phi(z_\alpha)}{q'(z_\alpha)}\left[\frac{\dot{\nu}_\gamma}{\sigma_\gamma} + q(z_\alpha)\frac{\dot{\sigma}_\gamma}{\sigma_\gamma}\right]. \qquad (3.28)$$

Pode agora substituir-se (3.28) na expressão da função diagnóstico, dada por (3.17),

$$D(z,\gamma) = \frac{m_\gamma[x_{\Phi(z),\gamma}]}{m_\gamma[x_{0.5,\gamma}]}\,\frac{\phi(0)}{\phi(z)},$$

isto é,

$$D(z,\gamma) = \frac{-\dfrac{\phi[z_{\Phi(z)}]}{q'[z_{\Phi(z)}]}\left[\dfrac{\dot{\nu}_\gamma}{\sigma_\gamma} + q[z_{\Phi(z)}]\dfrac{\dot{\sigma}_\gamma}{\sigma_\gamma}\right]}{-\dfrac{\phi(z_{0.5})}{q'(z_{0.5})}\left[\dfrac{\dot{\nu}_\gamma}{\sigma_\gamma} + q(z_{0.5})\dfrac{\dot{\sigma}_\gamma}{\sigma_\gamma}\right]}\,\frac{\phi(0)}{\phi(z)}\,.$$

Como $z_{\Phi(z)} = z$, por (3.22), e $z_{0.5} = 0$, fica,

$$D(z,\gamma) = \frac{-\dfrac{\phi(z)}{q'(z)}\left[\dfrac{\dot{\nu}_\gamma}{\sigma_\gamma} + q(z)\dfrac{\dot{\sigma}_\gamma}{\sigma_\gamma}\right]}{-\dfrac{\phi(0)}{q'(0)}\left[\dfrac{\dot{\nu}_\gamma}{\sigma_\gamma} + q(0)\dfrac{\dot{\sigma}_\gamma}{\sigma_\gamma}\right]}\,\frac{\phi(0)}{\phi(z)}\,.$$

Atendendo a que $q'(0) = 1$ e $q(0) = 0$, tem-se,

$$D(z,\gamma) = \frac{-\dfrac{\phi(z)}{q'(z)}\left[\dfrac{\dot{\nu}_\gamma}{\sigma_\gamma} + q(z)\dfrac{\dot{\sigma}_\gamma}{\sigma_\gamma}\right]}{-\phi(0)\dfrac{\dot{\nu}_\gamma}{\sigma_\gamma}}\,\frac{\phi(0)}{\phi(z)} =$$

$$= \frac{\dfrac{1}{q'(z)}\left[\dfrac{\dot{\nu}_\gamma}{\sigma_\gamma} + q(z)\dfrac{\dot{\sigma}_\gamma}{\sigma_\gamma}\right]}{\dfrac{\dot{\nu}_\gamma}{\sigma_\gamma}} = \frac{\dfrac{1}{q'(z)}\dfrac{1}{\sigma_\gamma}[\dot{\nu}_\gamma + q(z)\,\dot{\sigma}_\gamma]}{\dfrac{\dot{\nu}_\gamma}{\sigma_\gamma}} =$$

$$= \frac{\dot{\nu}_\gamma + q(z)\,\dot{\sigma}_\gamma}{q'(z)\,\dot{\nu}_\gamma} = \frac{1 + q(z)\dfrac{\dot{\sigma}_\gamma}{\dot{\nu}_\gamma}}{q'(z)} = \frac{1 + q(z)\,\varepsilon_\gamma}{q'(z)}\,, \quad \varepsilon_\gamma = \frac{\dot{\sigma}_\gamma}{\dot{\nu}_\gamma}\,,$$

como se queria demonstrar.

Com base na função diagnóstico e no Lema 1, pode analisar-se se uma dada família de distribuições é ou não uma GSTF.

No caso dos intervalos de confiança "Bootstrap", interessa particularmente trabalhar com famílias de transformações totalmente normalizáveis, as quais são caso particular das GSTF. Assim, diz-se que $\mathfrak{M}$ é uma "família de transformações normais escaladas", designada abreviadamente por NSTF (iniciais do inglês "normal scaled transformation family"), se existir uma função estritamente monótona, g, tal que,

$$\frac{g(X) - \nu_\gamma}{\sigma_\gamma} \sim Z, \tag{3.29}$$

onde ν_γ e σ_γ obedecem às condições definidas para (3.20).

Repare-se que as NSTF são caso particular das GSTF com $q(Z) = Z$. Para as NSTF, o Lema 1 apresenta a seguinte função diagnóstico,

$$D(z,\gamma) = 1 + z\,\varepsilon_\gamma\,. \tag{3.30}$$

Um caso particular das NSTF (logo, um caso ainda mais particular das GSTF) são as "famílias de transformações normais", designadas abreviadamente por NTF (iniciais do inglês "normal transformation family"). Diz-se que a família $\mathfrak{M}$ é uma NTF, se existir uma função estritamente monótona, g, tal que,

$$g(X) - \nu_\gamma \sim Z, \tag{3.31}$$

onde ν_γ obedece às condições definidas para (3.20).

Repare-se que as NTF são caso particular das NSTF com $\sigma_\gamma = 1$. Para as NTF, o Lema 1 apresenta a seguinte função diagnóstico,

$$D(z,\gamma) = 1. \tag{3.32}$$

As famílias NSTF e NTF são totalmente normalizáveis, através da função g.

O método dos percentis corrigido do enviesamento e o método dos percentis corrigido do enviesamento e da aceleração da variância pressupõem que a família de distribuições, a que pertence a "verdadeira" distribuição do estimador $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$, é (aproximadamente) uma NTF, no primeiro método, ou uma NSTF, no segundo método, o que vem constituir uma nítida generalização, quer em relação ao método tradicional, o qual utiliza a normalidade assintótica do estimador da máxima verosimilhança, quer em relação ao método dos percentis, o qual parte do princípio de que a família de distribuições a que pertence a "verdadeira" distribuição de $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$ é (aproximadamente) uma NTF muito especial, em que não se considera a possível existência de enviesamento.

A generalização e o cair de algumas hipóteses restritivas, proporcionadas pelo método dos percentis corrigido do enviesamento e pelo método dos percentis corrigido do enviesamento e da aceleração da variância, permitem-lhes um maior rigor e uma maior adequação à realidade (em especial por parte do último método), como vai ver-se adiante.

3.3.2 - O MÉTODO DOS PERCENTIS CORRIGIDO DO ENVIESAMENTO

Antes de entrar, propriamente, no método dos percentis corrigido do enviesamento, vai voltar-se ao método dos percentis e justificá-lo, com base na teoria da transformação.

O método dos percentis pressupõe que a família de distribuições a que pertence a "verdadeira" distribuição do estimador $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$ é (aproximadamente) uma NTF, em que não se considera a possível existência de enviesamento na transformação para a normalidade. Convém ver melhor este aspecto.

Considere-se uma função g, monótona crescente e defina-se,

$$\phi = g(\theta), \tag{3.33}$$

$$\hat{\phi} = g(\hat{\theta}), \tag{3.34}$$

$$\hat{\phi}^* = g(\hat{\theta}^*). \tag{3.35}$$

O método dos percentis parte do princípio que,

$$\frac{\hat{\phi} - \phi}{\tau} \sim N(0,1), \ \tau > 0, \tag{3.36}$$

e,

$$\frac{\hat{\phi}^* - \hat{\phi}}{\tau} \underset{*}{\sim} N(0,1), \ \tau > 0. \tag{3.37}$$

Repare-se que (3.36) não é mais do que (3.31), onde $\phi = \nu_y$. O facto de, em (3.31), se ter um desvio padrão unitário e, em (3.36), se ter um desvio padrão τ, não invalida a similitude entre as duas expressões - o facto relevante é que o desvio padrão, seja ele 1 ou τ, é constante.

De (3.36), pode tirar-se que, $\hat{\phi} \sim N(\phi,\tau^2)$, ou seja, $\hat{\phi}$ é um estimador centrado de ϕ, não se considerando nenhum enviesamento.

A expressão (3.36) permite, então, concluir que, para o método dos percentis, a família de distribuições de onde provém a "verdadeira" distribuição do estimador $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$ é (aproximadamente) caso particular de uma NTF em que se ignora o eventual enviesamento do estimador, admitindo-se, à partida, que é centrado.

As expressões (3.36) e (3.37) não devem ser entendidas de forma exacta, mas sim, como uma motivação para aplicar o método dos percentis. Por outras palavras, os fundamentos do método dos percentis não requerem que as expressões (3.36) e (3.37) se verifiquem exactamente, bastando apenas que se verifiquem aproximadamente - daí o dizer-se que, na óptica do método dos percentis, a família de distribuições a que pertence a "verdadeira" distribuição do estimador $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$ é aproximadamente (e não exactamente) uma NTF, em que não se considera a possível existência de enviesamento na estimação.

Por conveniência de exposição, vai pressupor-se que (3.36) e (3.37) se verificam exactamente, tendo-se, no entanto, subjacente que o raciocínio se mantem válido quando (3.36) e (3.37) se verificam apenas aproximadamente.

As expressões (3.36) e (3.37) significam que $\frac{\hat{\phi} - \phi}{\tau}$ é variável fulcral (pivô), no sentido em que a sua distribuição não depende de ϕ (logo não depende de θ), seja sob F (caso de 3.36) ou sob $\hat{F}$ (caso de 3.37), logo pode ser utilizada para construir um intervalo de confiança para θ. Veja-se como.

De (3.36), pode retirar-se,

$$P_F\left[z_\alpha \leq \frac{\hat{\phi} - \phi}{\tau} \leq z_{1-\alpha}\right] = 1-2\alpha\,. \tag{3.38}$$

Desenvolvendo (3.38), tem-se,

$$P_F\left[z_\alpha \leq \frac{\hat{\phi} - \phi}{\tau} \leq z_{1-\alpha}\right] = 1-2\alpha \Leftrightarrow$$

$$\Leftrightarrow P_F[z_\alpha\,\tau \leq \hat{\phi} - \phi \leq z_{1-\alpha}\,\tau] = 1-2\alpha \Leftrightarrow$$

$$\Leftrightarrow P_F[-\hat{\phi} + z_\alpha\,\tau \leq -\phi \leq -\hat{\phi} + z_{1-\alpha}\,\tau] = 1-2\alpha \Leftrightarrow$$

$$\Leftrightarrow P_F[\hat{\phi} - z_{1-\alpha}\,\tau \leq \phi \leq \hat{\phi} - z_\alpha\,\tau] = 1-2\alpha\,. \tag{3.39}$$

Se F fosse conhecida, podia propor-se,

$$[\hat{\phi} - z_{1-\alpha}\,\tau\ ;\ \hat{\phi} - z_\alpha\,\tau], \tag{3.40}$$

como intervalo de confiança a $(1-2\alpha)100\%$ para ϕ. No entanto, F é desconhecida, o que leva a ultrapassar o problema substituindo F por $\hat{F}$. Nesta linha de ideias, vão procurar-se as estimativas "Bootstrap" para os limites superior e inferior de (3.40).

Para o limite superior, e recorrendo a (3.37), tem-se,

$$P_{\hat{F}}[\hat{\phi}^* \leq \hat{\phi} - z_\alpha\,\tau] = P_{\hat{F}}\left[\frac{\hat{\phi}^* - \hat{\phi}}{\tau} \leq -z_\alpha\right] = \Phi(-z_\alpha) = 1-\alpha\,.$$

Sendo $\hat{H}(s) = \dfrac{\#\{\hat{\phi}^{*b} \leq s\}}{B}$, $\hat{\phi}^{*b} = g(\hat{\theta}^{*b})$, a função de distribuição "Bootstrap" de $\hat{\phi}$, tem-se que $P_{\hat{F}}[\hat{\phi}^* \leq \hat{\phi} - z_\alpha \tau] = \hat{H}(\hat{\phi} - z_\alpha \tau)$, se $B \to +\infty$. Nestas condições,

$$\hat{H}(\hat{\phi} - z_\alpha \tau) = 1-\alpha \Leftrightarrow \hat{\phi} - z_\alpha \tau = \hat{H}^{-1}(1-\alpha). \qquad (3.41)$$

Para o limite inferior, e tornando a recorrer a (3.37), tem-se,

$$P_{\hat{F}}[\hat{\phi}^* > \hat{\phi} - z_{1-\alpha} \tau] = 1 - P_{\hat{F}}[\hat{\phi}^* \leq \hat{\phi} - z_{1-\alpha} \tau] =$$

$$= 1 - P_{\hat{F}}\left[\frac{\hat{\phi}^* - \hat{\phi}}{\tau} \leq -z_{1-\alpha}\right] = 1 - \Phi(-z_{1-\alpha}) = 1-\alpha .$$

Como $1 - P_{\hat{F}}[\hat{\phi}^* \leq \hat{\phi} - z_{1-\alpha} \tau] = 1 - \hat{H}(\hat{\phi} - z_{1-\alpha} \tau)$, supondo que $B \to +\infty$, tem-se,

$$1 - \hat{H}(\hat{\phi} - z_{1-\alpha} \tau) = 1-\alpha \Leftrightarrow \hat{\phi} - z_{1-\alpha} \tau = \hat{H}^{-1}(\alpha). \qquad (3.42)$$

Então, um intervalo de confiança "Bootstrap" para ϕ, a $(1-2\alpha)100\%$, é dado por,

$$\left[\hat{H}^{-1}(\alpha) ; \hat{H}^{-1}(1-\alpha)\right]. \qquad (3.43)$$

Note-se que a igualdade $\hat{H}(s) = P_{\hat{F}}(\hat{\phi}^* \leq s) \equiv P_*(\hat{\phi}^* \leq s)$ só é válida quando $B \to +\infty$, ou seja, $\hat{H}(s)$ não é a verdadeira função de distribuição "Bootstrap" de $\hat{\phi}$, mas apenas uma sua estimativa, resultante do já conhecido algoritmo de Monte Carlo (com base na construção de B amostras "Bootstrap"). Nestas condições, tenha-se em atenção que o nível de confiança do intervalo (3.43) é aproximado e não exacto.

Falta agora passar de um intervalo de confiança em ϕ para um intervalo de confiança em θ. De (3.33) tira-se $\theta = g^{-1}(\phi)$, logo aplicando a função inversa g^{-1} aos limites do intervalo (3.43), vem,

$$g^{-1}[\hat{H}^{-1}(\alpha)] = g^{-1}\left\{g[\hat{G}^{-1}(\alpha)]\right\} = \hat{G}^{-1}(\alpha), \tag{3.44}$$

$$g^{-1}[\hat{H}^{-1}(1-\alpha)] = g^{-1}\left\{g[\hat{G}^{-1}(1-\alpha)]\right\} = \hat{G}^{-1}(1-\alpha). \tag{3.45}$$

Os resultados (3.44) e (3.45) assentam na relação,

$$\hat{H}(s) = \hat{G}[g^{-1}(s)], \tag{3.46}$$

da qual sai,

$$\hat{H}^{-1}(s) = \{\hat{G}[g^{-1}(s)]\}^{-1} = (g^{-1})^{-1}[\hat{G}^{-1}(s)] = g[\hat{G}^{-1}(s)]. \tag{3.47}$$

Então, de (3.44) e de (3.45), pode construir-se o intervalo de confiança a $(1-2\alpha)100\%$ (aproximadamente) para θ, dado pelo método dos percentis,

$$\left[\hat{G}^{-1}(\alpha)\ ;\ \hat{G}^{-1}(1-\alpha)\right] = \left[\hat{\theta}_{PER}(LI)\ ;\ \hat{\theta}_{PER}(LS)\right],$$

o qual é, evidentemente, igual ao deduzido no ponto 3.2.

Se as expressões (3.36) e (3.37) forem exactas, então o intervalo de confiança dado pelo método dos percentis coincide com o intervalo exacto (3.1) (esquecendo que B assume, na prática, um valor finito). Caso contrário, o intervalo do método dos percentis é aproximado e tem-se que as diferenças $\left|\hat{\theta}_{PER}(LI) - \theta_{LI}\right|$ e $\left|\hat{\theta}_{PER}(LS) - \theta_{LS}\right|$ são $O_p(n^{-1})$ (pelo menos em domínios paramétricos unidimensionais, podendo-se conjecturar que também o sejam em domínios não paramétricos), tal como já se tinha referido no último parágrafo do ponto 3.2.

Fica assim justificado o método dos percentis, com base na teoria da transformação, o que vai permitir fazer uma ponte de ligação entre este método e o método dos percentis corrigido do enviesamento. Na verdade, a grande diferença entre estes dois métodos reside no facto de o último supor que a família de distribuições a que pertence a "verdadeira" distribuição do estimador $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$ é (aproximadamente) uma NTF, considerando-se a possível existência de enviesamento na estimação. Assim, mantêm-se válidas as expressões (3.33), (3.34) e (3.35), mas as variáveis aleatórias presentes em (3.36) e (3.37) são substituídas por,

$$\frac{\hat{\phi} - \phi}{\tau} \sim N(-z_0, 1), \ \tau > 0, \tag{3.48}$$

e,

$$\frac{\hat{\phi}^* - \hat{\phi}}{\tau} \underset{*}{\sim} N(-z_0, 1), \ \tau > 0. \tag{3.49}$$

Repare-se que (3.48) não é mais do que (3.31), onde $v_\gamma = \phi - z_0 \tau$. Como foi dito, o facto de, em (3.31), se ter um desvio padrão unitário e, em (3.48), se ter um desvio padrão τ não invalida a similitude entre as duas expressões.

As expressões (3.48) e (3.49) não devem ser entendidas de forma exacta, mas sim como uma motivação para aplicar o método dos percentis corrigido do enviesamento. Por outras palavras, os fundamentos do método dos percentis corrigido do enviesamento não requerem que as expressões (3.48) e (3.49) se verifiquem exactamente, bastando apenas que se verifiquem aproximadamente - daí o dizer-se que, na óptica do método dos percentis corrigido do enviesamento, a família de distribuições a que pertence a "verdadeira" distribuição do estimador $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$ é aproximadamente

(e não exactamente) uma NTF, em que se considera a possível existência de enviesamento na estimação.

Por conveniência de exposição, vai pressupor-se que (3.48) e (3.49) se verificam exactamente, tendo-se, no entanto, subjacente que o raciocínio se mantem válido quando (3.48) e (3.49) se verificam apenas aproximadamente.

Veja-se como se determina z_0, a chamada constante de enviesamento, porque mede o enviesamento de $\hat{\phi}$ como estimador de ϕ, [$E(\hat{\phi}) - \phi = -z_0 \tau$].

De (3.48), pode concluir-se que,

$$\frac{\hat{\phi} - \phi}{\tau} + z_0 \sim N(0,1) \Leftrightarrow$$

$$\Leftrightarrow \frac{\hat{\phi} - (\phi - z_0 \tau)}{\tau} \sim N(0,1) \Leftrightarrow$$

$$\Leftrightarrow \hat{\phi} = (\phi - z_0 \tau) + \tau Z, \quad Z \sim N(0,1).$$

Então,

$$P_F(\hat{\phi} \leq \phi) = P_F[(\phi - z_0 \tau) + \tau Z \leq \phi] =$$

$$= P_F[\tau (Z - z_0) \leq 0] = P_F[Z \leq z_0] = \Phi(z_0). \qquad (3.50)$$

Atendendo a (3.33), (3.34) e (3.50), vem,

$$P_F(\hat{\theta} \leq \theta) = P_F[g(\hat{\theta}) \leq g(\theta)] = P_F(\hat{\phi} \leq \phi) = \Phi(z_0). \qquad (3.51)$$

Como (3.48) e (3.49) têm a mesma distribuição,

$$P_F(\hat{\phi} \leq \phi) = P_{\hat{F}}(\hat{\phi}^* \leq \hat{\phi}) = \Phi(z_0). \qquad (3.52)$$

Como $\hat{G}(\hat{\theta}) = P_{\hat{F}}(\hat{\theta}^* \leq \hat{\theta})$, quando $B \to +\infty$, pode introduzir-se este resultado em (3.52), ficando,

$$\hat{G}(\hat{\theta}) = \Phi(z_0) \Leftrightarrow z_0 = \Phi^{-1}[\hat{G}(\hat{\theta})]. \qquad (3.53)$$

Repare-se que a obtenção de z_0 , através de (3.53), não é rigorosamente exacta, na medida em que $\hat{G}(t)$ não é exactamente a função de distribuição "Bootstrap" de $\hat{\theta}$, mas apenas uma sua estimativa.

Com as hipóteses do método dos percentis corrigido do enviesamento, a variável fulcral é $\dfrac{\hat{\phi} - (\phi - z_0\,\tau)}{\tau} \sim N(0,1)$, no sentido em que a sua distribuição não depende de ϕ (logo não depende de θ), seja sob F (caso de (3.48)) ou sob $\hat{F}$ (caso de (3.49)), logo pode ser utilizada para construir um intervalo de confiança para θ. Vai construir-se esse intervalo.

$$P_F\left[z_\alpha \leq \frac{\hat{\phi} - (\phi - z_0\,\tau)}{\tau} \leq z_{1-\alpha}\right] = 1-2\alpha \Leftrightarrow$$

$$\Leftrightarrow P_F\left[z_\alpha\,\tau \leq \hat{\phi} - (\phi - z_0\,\tau) \leq z_{1-\alpha}\,\tau\right] = 1-2\alpha \Leftrightarrow$$

$$\Leftrightarrow P_F\left[-\hat{\phi} - z_0\,\tau + z_\alpha\,\tau \leq -\phi \leq -\hat{\phi} - z_0\,\tau + z_{1-\alpha}\,\tau\right] = 1-2\alpha \Leftrightarrow$$

$$\Leftrightarrow P_F\left[\hat{\phi} + z_0\,\tau - z_{1-\alpha}\,\tau \leq \phi \leq \hat{\phi} + z_0\,\tau - z_\alpha\,\tau\right] = 1-2\alpha\,. \qquad (3.54)$$

Se F fosse conhecida, poderia propor-se,

$$\left[\hat{\phi} + z_0\,\tau - z_{1-\alpha}\,\tau \;;\; \hat{\phi} + z_0\,\tau - z_\alpha\,\tau\right], \qquad (3.55)$$

como intervalo de confiança a $(1-2\alpha)100\%$ para ϕ. No entanto, F é desconhecida, o que leva a ultrapassar o problema substituindo F por $\hat{F}$. Nesta linha de ideias, vão procurar-se as estimativas "Bootstrap" para os limites superior e inferior de (3.55).

Para o limite superior, e recorrendo a (3.49), tem-se,

$$P_{\hat{F}}[\hat{\phi}^* \le \hat{\phi} + z_0\,\tau - z_\alpha\,\tau] = P_{\hat{F}}[\hat{\phi}^* - \hat{\phi} + z_0\,\tau \le 2\,z_0\,\tau - z_\alpha\,\tau] =$$

$$= P_{\hat{F}}\left[\frac{\hat{\phi}^* - \hat{\phi} + z_0\,\tau}{\tau} \le 2\,z_0 - z_\alpha\right] = \Phi(2\,z_0 - z_\alpha).$$

Como $P_{\hat{F}}[\hat{\phi}^* \le \hat{\phi} + z_0\,\tau - z_\alpha\,\tau] = \hat{H}(\hat{\phi} + z_0\,\tau - z_\alpha\,\tau)$, supondo que $B \to +\infty$, tem-se,

$$\hat{H}(\hat{\phi} + z_0\,\tau - z_\alpha\,\tau) = \Phi(2\,z_0 - z_\alpha) \Leftrightarrow$$

$$\Leftrightarrow \hat{\phi} + z_0\,\tau - z_\alpha\,\tau = \hat{H}^{-1}[\Phi(2\,z_0 - z_\alpha)]. \qquad (3.56)$$

Para o limite inferior, e tornando a recorrer a (3.49), tem-se,

$$P_{\hat{F}}[\hat{\phi}^* > \hat{\phi} + z_0\,\tau - z_{1-\alpha}\,\tau] = 1 - P_{\hat{F}}[\hat{\phi}^* \le \hat{\phi} + z_0\,\tau - z_{1-\alpha}\,\tau] =$$

$$= 1 - P_{\hat{F}}[\hat{\phi}^* - \hat{\phi} + z_0\,\tau \le 2\,z_0\,\tau - z_{1-\alpha}\,\tau] =$$

$$= 1 - P_{\hat{F}}\left[\frac{\hat{\phi}^* - \hat{\phi} + z_0\,\tau}{\tau} \le 2\,z_0 - z_{1-\alpha}\right] = 1 - \Phi(2\,z_0 - z_{1-\alpha}).$$

Como $1 - P_{\hat{F}}[\hat{\phi}^* \le \hat{\phi} + z_0\,\tau - z_{1-\alpha}\,\tau] = 1 - \hat{H}(\hat{\phi} + z_0\,\tau - z_{1-\alpha}\,\tau)$, supondo que $B \to +\infty$, tem-se,

$$1 - \hat{H}(\hat{\phi} + z_0\,\tau - z_{1-\alpha}\,\tau) = 1 - \Phi(2\,z_0 - z_{1-\alpha}) \Leftrightarrow$$

$$\Leftrightarrow \hat{\phi} + z_0\,\tau - z_{1-\alpha}\,\tau = \hat{H}^{-1}[\Phi(2\,z_0 - z_{1-\alpha})]. \qquad (3.57)$$

Então, um intervalo de confiança "Bootstrap" para ϕ, a $(1-2\alpha)100\%$, é dado por,

$$\left[\hat{H}^{-1}[\Phi(2\,z_0 - z_{1-\alpha})]\ ;\ \hat{H}^{-1}[\Phi(2\,z_0 - z_\alpha)]\right]. \qquad (3.58)$$

Como a igualdade $\hat{H}(s) = P_{\hat{F}}(\hat{\phi}^* \leq s) \equiv P_*(\hat{\phi}^* \leq s)$ só é válida quando $B \to +\infty$, ou seja, como $\hat{H}(s)$ não é a verdadeira função de distribuição "Bootstrap" de $\hat{\phi}$, mas apenas uma sua estimativa, o nível de confiança do intervalo (3.58) é aproximado e não exacto.

Falta agora passar de um intervalo de confiança em ϕ para um intervalo de confiança em θ. De (3.33), tira-se $\theta = g^{-1}(\phi)$, logo vai aplicar-se a função inversa g^{-1} aos limites do intervalo (3.58),

$$\begin{aligned} & g^{-1}\left\{\hat{H}^{-1}[\Phi(2\,z_0 - z_{1-\alpha})]\right\} = \\ & = g^{-1}\left\{g(\hat{G}^{-1}[\Phi(2\,z_0 - z_{1-\alpha})])\right\} = \\ & = \hat{G}^{-1}[\Phi(2\,z_0 - z_{1-\alpha})], \end{aligned} \tag{3.59}$$

e,

$$\begin{aligned} & g^{-1}\left\{\hat{H}^{-1}[\Phi(2\,z_0 - z_{\alpha})]\right\} = \\ & = g^{-1}\left\{g(\hat{G}^{-1}[\Phi(2\,z_0 - z_{\alpha})])\right\} = \\ & = \hat{G}^{-1}[\Phi(2\,z_0 - z_{\alpha})]. \end{aligned} \tag{3.60}$$

Os resultados (3.59) e (3.60) assentam nas relações (3.46) e (3.47).

Então, de (3.59) e de (3.60), pode construir-se o intervalo de confiança a $(1-2\alpha)100\%$ (aproximadamente) para θ, dado pelo método dos percentis corrigido do enviesamento,

$$\left[\hat{\theta}_{BC}(LI)\ ;\ \hat{\theta}_{BC}(LS)\right] = \left[\hat{G}^{-1}[\Phi(2\,z_0 - z_{1-\alpha})]\ ;\ \hat{G}^{-1}[\Phi(2\,z_0 - z_{\alpha})]\right]. \tag{3.61}$$

Note-se que,

$$\hat{\theta}_{BC}(LI) = t_0 \ : \ t_0 = \inf_t \frac{\#\{\hat{\theta}^{*b} \leq t\}}{B} \geq \Phi(2\, z_0 - z_{1-\alpha}), \qquad (3.62)$$

e,

$$\hat{\theta}_{BC}(LS) = t_1 \ : \ t_1 = \inf_t \frac{\#\{\hat{\theta}^{*b} \leq t\}}{B} \geq \Phi(2\, z_0 - z_{\alpha}). \qquad (3.63)$$

Se não houver enviesamento, isto é, se $z_0 = 0$, então $\hat{\theta}_{BC}(LI) = \hat{\theta}_{PER}(LI)$ e $\hat{\theta}_{BC}(LS) = \hat{\theta}_{PER}(LS)$, ou seja, o método dos percentis corrigido do enviesamento conduz ao mesmo resultado que o método dos percentis.

Se as expressões (3.48) e (3.49) forem exactas, então o intervalo de confiança dado pelo método dos percentis corrigido do enviesamento coincide com o intervalo exacto (3.1) (esquecendo que B assume, na prática, um valor finito). Caso contrário, o intervalo (3.61) é aproximado e tem-se que as diferenças $|\hat{\theta}_{BC}(LI) - \theta_{LI}|$ e $|\hat{\theta}_{BC}(LS) - \theta_{LS}|$ são $O_P(n^{-1})$ (pelo menos em domínios paramétricos unidimensionais, podendo-se conjecturar que também o sejam em domínios não paramétricos) [veja-se Dicicio e Romano (1988) - pg. 340], pelo que, para pequenas amostras, (3.61) pode ser significativamente diferente de (3.1) que é o intervalo de confiança exacto, a $(1-2\alpha)100\%$, para θ. Daí que Efron tenha pesquisado outros métodos "Bootstrap" para construção de intervalos de confiança, tendo chegado ao método dos percentis corrigido do enviesamento e da aceleração da variância [veja-se Efron (1987)].

3.3.3 - O MÉTODO DOS PERCENTIS CORRIGIDO DO ENVIESAMENTO E DA ACELERAÇÃO DA VARIÂNCIA

O método dos percentis corrigido do enviesamento e da aceleração da variância é mais geral do que os métodos atrás apresentados, na medida em que parte do pressuposto de que a família de distribuições, à qual pertence a "verdadeira" distribuição do estimador $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$ é (aproximadamente) uma NSTF.

Os intervalos de confiança deduzidos por Efron baseiam-se no princípio de que a família de distribuições a que pertence a "verdadeira" distribuição do estimador $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$ é normalizável. No entanto, o raciocínio pode ser facilmente generalizado, para os casos em que a citada família de distribuições não é normalizável, mas pode transformar-se, através de uma função monótona crescente g, numa família de distribuições simétricas conhecidas [veja-se Murteira (1988b) - pg. 9].

Assim, mantêm-se válidas as expressões (3.33), (3.34) e (3.35), mas as variáveis aleatórias presentes em (3.48) e (3.49) são substituídas por outras hipóteses,

$$\frac{\hat{\phi} - \phi}{\tau} \sim N(-z_0 \sigma_\phi, \sigma_\phi^2), \ \tau > 0, \tag{3.64}$$

e,

$$\frac{\hat{\phi}^* - \hat{\phi}}{\tau} \sim N(-z_0 \sigma_{\hat{\phi}}, \sigma_{\hat{\phi}}^2), \ \tau > 0, \tag{3.65}$$

onde,

$$\sigma_{\phi} = 1 + a\,\phi > 0 \quad e \quad \sigma_{\hat{\phi}} = 1 + a\,\hat{\phi} > 0, \tag{3.66}$$

sendo "a" a chamada constante de aceleração.

Pode obter-se uma interpretação da constante de aceleração através das expressões $\sigma_{\phi} = 1 + a\,\phi$ e $\sigma_{\phi_0} = 1 + a\,\phi_0$. Subtraindo ordenadamente ambos os membros destas equações, tem-se:

$$\sigma_{\phi} - \sigma_{\phi_0} = a\,(\phi - \phi_0) \Leftrightarrow \sigma_{\phi} = \sigma_{\phi_0} + a\,(\phi - \phi_0) \Leftrightarrow$$

$$\Leftrightarrow \sigma_{\phi} = \sigma_{\phi_0}\left[1 + a\,\frac{(\phi - \phi_0)}{\sigma_{\phi_0}}\right] \Leftrightarrow \frac{\sigma_{\phi}}{\sigma_{\phi_0}} = 1 + a\,\frac{(\phi - \phi_0)}{\sigma_{\phi_0}}\,.$$

Deste último resultado, sai,

$$a = \frac{d\left(\frac{\sigma_{\phi}}{\sigma_{\phi_0}}\right)}{d\left(\frac{\phi - \phi_0}{\sigma_{\phi_0}}\right)}\,,$$

ou seja, a constante de aceleração mede a variação relativa de σ_{ϕ}, por cada cada unidade de variação de ϕ referida ao desvio padrão inicial σ_{ϕ_0} [veja-se Efron (1987) - pg. 175].

Repare-se que (3.64) não é mais do que (3.29), onde $v_{\gamma} = \phi - z_0\,\sigma_{\phi}\,\tau$ e $\sigma_{\gamma} = \tau\,\sigma_{\phi}$.

As expressões (3.64) e (3.65) não devem ser entendidas de forma exacta, mas sim, como uma motivação para aplicar o método dos percentis corrigido do enviesamento e da aceleração da variância. Por outras palavras, os fundamentos deste método não requerem que as expressões (3.64) e (3.65) se

verifiquem exactamente, bastando apenas que se verifiquem aproximadamente - daí o dizer-se que, na óptica do método dos percentis corrigido do enviesamento e da aceleração da variância, a família de distribuições a que pertence a "verdadeira" distribuição do estimador $\hat{\theta} \equiv \hat{\theta}(X_1, X_2, ..., X_n)$ é aproximadamente (e não exactamente) uma NSTF.

Por conveniência de exposição, vai pressupor-se que (3.64) e (3.65) se verificam exactamente, tendo-se, no entanto, subjacente que o raciocínio se mantem válido quando (3.64) e (3.65) se verificam apenas aproximadamente.

Com estas hipóteses, a variável fulcral do método dos percentis corrigido do enviesamento e da aceleração da variância é $\dfrac{\hat{\phi} - (\phi - z_0\,\sigma_\phi\,\tau)}{\tau\,\sigma_\phi} \sim N(0,1)$, no sentido em que a sua distribuição não depende de ϕ (logo não depende de θ), seja sob F (caso de (3.64)) ou sob $\hat{F}$ (caso de (3.65)).

Este método vem abrir a possibilidade de o desvio padrão do estimador não ser constante, o que alarga a classe de problemas a que o método se pode aplicar com resultados satisfatórios. Na verdade, as hipóteses (3.36) e (3.37) (do método dos percentis) e (3.48) e (3.49) (do método dos percentis corrigido do enviesamento) impõem, simultâneamente, a normalização e a estabilização da variância, objectivos estes que são parcialmente antagónicos nas famílias de distribuições mais vulgares [veja-se Efron (1987) - pg. 172], o que se pode traduzir (negativamente) na qualidade dos resultados obtidos por estes métodos. Aliás, Schenker tinha alertado para a existência de algumas deficiências nos métodos dos percentis e dos percentis corrigido do enviesamento (os quais impõem um desvio padrão constante), nomeadamente, ao nível da probabilidade de cobertura (nível de confiança real) dos respectivos intervalos de confiança "Bootstrap", a qual tende a situar-se

significativamente abaixo do nível de confiança $(1-2\alpha)100\%$ pretendido, para amostras não muito grandes e para alguns problemas concretos estudados [veja-se Schenker (1985)]. O método dos percentis corrigido do enviesamento e da aceleração da variância é uma resposta a estas deficiências, na medida em que se baseia em hipóteses não tão restritivas (cai a hipótese de desvio padrão constante).

Para evitar expressões analíticas muito "pesadas", a dedução dos intervalos de confiança "Bootstrap" vai ter por base, não exactamente as hipóteses (3.64) e (3.65), mas outras hipóteses similares, em que se faz $\tau = 1$,

$$\hat{\phi} - \phi \sim N(-z_0 \sigma_\phi, \sigma_\phi^2), \tag{3.67}$$

e,

$$\hat{\phi}^* - \hat{\phi} \underset{*}{\sim} N(-z_0 \sigma_{\hat{\phi}}, \sigma_{\hat{\phi}}^2). \tag{3.68}$$

Os dois pares de expressões - (3.64) e (3.65), por um lado, (3.67) e (3.68) por outro lado - são equivalentes, na medida em que o primeiro pode sempre reduzir-se ao segundo. Para demonstrar esta asserção, tem de provar-se que (3.64) é equivalente a (3.67) e que (3.65) é equivalente a (3.68). Vai apenas provar-se que (3.64) é equivalente a (3.67). Parte-se do caso mais geral, ou seja, de (3.64), o qual permite concluir que,

$$\frac{\hat{\phi} - (\phi - z_0 \sigma_\phi \tau)}{\tau \sigma_\phi} \sim N(0,1).$$

Como $Z \sim N(0,1)$, vem,

$$\frac{\hat{\phi} - (\phi - z_0 \sigma_\phi \tau)}{\tau \sigma_\phi} = Z \Leftrightarrow \hat{\phi} = Z \tau \sigma_\phi + \phi - z_0 \sigma_\phi \tau . \tag{3.69}$$

Considerando as transformações $\phi' = \phi/\tau$, $\hat{\phi}' = \hat{\phi}/\tau$ e desenvolvendo a partir de (3.69), tem-se,

$$\frac{\hat{\phi}}{\tau} = Z\,\sigma_{\phi} + \frac{\phi}{\tau} - z_0\,\sigma_{\phi} \Leftrightarrow \hat{\phi}' = \phi' + \sigma_{\phi}\,(Z - z_0). \qquad (3.70)$$

Como,

$$\sigma_{\phi} = 1 + a\,\phi = 1 + a\,\tau\,\frac{\phi}{\tau} = 1 + a'\,\phi' = \sigma_{\phi'}\,,\ a' = a\,\tau\,,$$

pode substituir-se σ_{ϕ} por $\sigma_{\phi'}$, em (3.70),

$$\hat{\phi}' = \phi' + \sigma_{\phi'}\,(Z - z_0) \Leftrightarrow$$

$$\Leftrightarrow \frac{\hat{\phi}' - \phi'}{\sigma_{\phi'}} = Z - z_0 \Leftrightarrow$$

$$\Leftrightarrow \frac{\hat{\phi}' - (\phi' - z_0\,\sigma_{\phi'})}{\sigma_{\phi'}} = Z. \qquad (3.71)$$

De (3.71), conclui-se que,

$$\frac{\hat{\phi}' - (\phi' - z_0\,\sigma_{\phi'})}{\sigma_{\phi'}} \sim N(0,1) \Leftrightarrow \hat{\phi}' - \phi' \sim N(-z_0\,\sigma_{\phi'}, \sigma^2_{\phi'}),$$

resultado este que é precisamente (3.67). Assim se prova que (3.64) é equivalente a (3.67) , bastando, para tal, escolher uma transformação g adequada que transforme θ (e $\hat{\theta}$) não em ϕ (e em $\hat{\phi}$), mas sim em ϕ' (e em $\hat{\phi}'$).

De forma idêntica, se demonstrava a equivalência entre (3.65) e (3.68).

Para passar à construção dos intervalos de confiança "Bootstrap" corrigidos do enviesamento e da aceleração da variância, com base nas

hipóteses (3.66), (3.67) e (3.68), note-se que, da expressão (3.67), se pode deduzir,

$$\hat{\phi} = \phi + \sigma_{\phi} (Z - z_0), \tag{3.72}$$

porque $Z \sim N(0,1)$, donde sai que,

$$\hat{\phi} - \phi = \sigma_{\phi} (Z - z_0) \sim N(-z_0 \sigma_{\phi}, \sigma_{\phi}^2),$$

resultado este idêntico a (3.67).

Utilizando (3.66) e (3.72) chega-se à conclusão de que,

$$1 + a \hat{\phi} = [1 + a \phi] [1 + a (Z - z_0)], \tag{3.73}$$

como se pode ver,

$$[1 + a \phi] [1 + a (Z - z_0)] = 1 + a (Z - z_0) + a \phi [1 + a (Z - z_0)] =$$

$$= 1 + a (Z - z_0) + a \phi + a^2 \phi (Z - z_0) = 1 + a \phi + a (Z - z_0) (1 + a \phi) =$$

$$= 1 + a \phi + a (Z - z_0) \sigma_{\phi} = 1 + a [\phi + (Z - z_0) \sigma_{\phi}] = 1 + a \hat{\phi}.$$

Tomando logaritmos neperianos em ambos os membros de (3.73), tem-se,

$$\ln (1 + a \hat{\phi}) = \ln \{[1 + a \phi] [1 + a (Z - z_0)]\} \Leftrightarrow$$

$$\Leftrightarrow \ln (1 + a \hat{\phi}) = \ln (1 + a \phi) + \ln [1 + a (Z - z_0)].$$

Fazendo,

$$\hat{\xi} = \ln (1 + a \hat{\phi}), \tag{3.74}$$

$$\xi = \ln (1 + a \phi), \tag{3.75}$$

e,

$$W = \ln [1 + a (Z - z_0)], \tag{3.76}$$

tem-se,

$$\hat{\xi} = \xi + W \Leftrightarrow W = \hat{\xi} - \xi . \tag{3.77}$$

Para as expressões (3.74), (3.75) e (3.76) serem válidas, assume-se, sem perda de generalidade, que $1 + a\,\hat{\phi} > 0$, $1 + a\,\phi > 0$ e $1 + a\,(Z - z_0) > 0$ [veja-se Efron (1987) - pg. 174].

De (3.77) sai o intervalo de confiança central a $(1-2\alpha)100\%$ para ξ,

$$[\hat{\xi} - w_{1-\alpha} \; ; \; \hat{\xi} - w_{\alpha}] \tag{3.78}$$

onde,

$$w_{1-\alpha} : P(W \leq w_{1-\alpha}) = 1 - \alpha \quad e \quad w_{\alpha} : P(W \leq w_{\alpha}) = \alpha .$$

O intervalo (3.78) deduz-se, a partir de (3.77), da seguinte forma,

$$P(w_{\alpha} \leq W \leq w_{1-\alpha}) = 1 - 2\,\alpha \Leftrightarrow$$

$$\Leftrightarrow P(w_{\alpha} \leq \hat{\xi} - \xi \leq w_{1-\alpha}) = 1 - 2\,\alpha \Leftrightarrow$$

$$\Leftrightarrow P(-\hat{\xi} + w_{\alpha} \leq -\xi \leq -\hat{\xi} + w_{1-\alpha}) = 1 - 2\,\alpha \Leftrightarrow$$

$$\Leftrightarrow P(\hat{\xi} - w_{1-\alpha} \leq \xi \leq \hat{\xi} - w_{\alpha}) = 1 - 2\,\alpha .$$

Para transformar o intervalo (3.78), da escala dos ξ para a escala dos ϕ, ou seja, passar de $[\xi(LI) \; ; \; \xi(LS)]$ para $[\phi(LI) \; ; \; \phi(LS)]$, há que atender às relações (3.74), (3.75) e (3.76), as quais permitem deduzir,

$$\phi = \frac{e^{\ln(1 + a\phi)} - 1}{a} = \frac{e^{\xi} - 1}{a}, \tag{3.79}$$

$$e^{\hat{\xi}} = 1 + a\,\hat{\phi}, \tag{3.80}$$

$$e^{W} = 1 + a\,(Z - z_0). \tag{3.81}$$

Consequentemente, utilizando (3.79), (3.80) e (3.81), e relembrando que $z_{1-\alpha} = -z_{\alpha}$, devido à simetria da normal estandardizada, tem-se,

$$\phi(LI) = \frac{e^{\xi(LI)} - 1}{a} = \frac{e^{\hat{\xi} - w_{1-\alpha}} - 1}{a} = \frac{e^{\hat{\xi}}\, e^{-w_{1-\alpha}} - 1}{a} =$$

$$= \frac{\dfrac{1 + a\,\hat{\phi}}{1 + a\,(z_{1-\alpha} - z_0)} - 1}{a} = \frac{1 + a\,\hat{\phi}}{a\,[1 + a\,(z_{1-\alpha} - z_0)]} - \frac{1}{a} =$$

$$= \frac{1 + a\,\hat{\phi} - [1 + a\,(z_{1-\alpha} - z_0)]}{a\,[1 + a\,(z_{1-\alpha} - z_0)]} = \frac{\hat{\phi} - (z_{1-\alpha} - z_0)}{1 + a\,(z_{1-\alpha} - z_0)} =$$

$$= \frac{\hat{\phi} - (-z_{\alpha} - z_0)}{1 + a\,(-z_{\alpha} - z_0)} = \frac{\hat{\phi} + (z_{\alpha} + z_0)}{1 - a\,(z_{\alpha} + z_0)} =$$

$$= \hat{\phi} + \frac{\hat{\phi} + (z_{\alpha} + z_0)}{1 - a\,(z_{\alpha} + z_0)} - \hat{\phi} =$$

$$= \hat{\phi} + \frac{\hat{\phi} + (z_{\alpha} + z_0) - \hat{\phi}\,[1 - a\,(z_{\alpha} + z_0)]}{1 - a\,(z_{\alpha} + z_0)} =$$

$$= \hat{\phi} + \frac{(z_{\alpha} + z_0) + \hat{\phi}\,a\,(z_{\alpha} + z_0)]}{1 - a\,(z_{\alpha} + z_0)} = \hat{\phi} + \frac{(z_0 + z_{\alpha})\,(1 + a\,\hat{\phi})}{1 - a\,(z_0 + z_{\alpha})} =$$

$$= \hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_{\alpha}}{1 - a\,(z_0 + z_{\alpha})}\,. \tag{3.82}$$

Por processo idêntico, consegue provar-se que,

$$\phi(LS) = \hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\,. \tag{3.83}$$

De (3.82) e (3.83), sai o intervalo de confiança a $(1-2\alpha)100\%$ para ϕ,

$$[\phi(LI)\,;\,\phi(LS)] = \left[\hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\,;\, \hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\right]. \quad (3.84)$$

Se F fosse conhecida, bastaria aplicar a transformação inversa g^{-1} aos limites inferior e superior de (3.84), para se obter o intervalo de confiança a $(1-2\alpha)100\%$ para θ. No entanto, F é desconhecida, o que leva a ultrapassar o problema substituindo F por $\hat{F}$. Nesta linha de ideias, vão procurar-se as estimativas "Bootstrap" para os limites inferior e superior de (3.84).

Para o limite superior, e recorrendo a (3.68), tem-se,

$$P_{\hat{F}}\left[\hat{\phi}^* \le \hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\right] =$$

$$= P_{\hat{F}}\left[\frac{\hat{\phi}^* - \hat{\phi} + z_0\,\sigma_{\hat{\phi}}}{\sigma_{\hat{\phi}}} \le \left(\sigma_{\hat{\phi}} \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})} + z_0\,\sigma_{\hat{\phi}}\right) \frac{1}{\sigma_{\hat{\phi}}}\right] =$$

$$= P_{\hat{F}}\left[\frac{\hat{\phi}^* - \hat{\phi} + z_0\,\sigma_{\hat{\phi}}}{\sigma_{\hat{\phi}}} \le z_0 + \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\right] =$$

$$= \Phi\left[z_0 + \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\right].$$

Como $P_{\hat{F}}\left[\hat{\phi}^* \le \hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\right] = \hat{H}\left[\hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\right]$,

quando $B \to +\infty$, fica,

$$\hat{H}\left[\hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\right] = \Phi\left[z_0 + \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\right] \Leftrightarrow$$

$$\Leftrightarrow \hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})} = \hat{H}^{-1}\left\{\Phi\left[z_0 + \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\right]\right\}. \quad (3.85)$$

Para o limite inferior, e tornando a recorrer a (3.68), tem-se,

$$P_{\hat{F}}\left[\hat{\phi}^* > \hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\right] =$$

$$= 1 - P_{\hat{F}}\left[\hat{\phi}^* \leq \hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\right] =$$

$$= 1 - P_{\hat{F}}\left[\frac{\hat{\phi}^* - \hat{\phi} + z_0\,\sigma_{\hat{\phi}}}{\sigma_{\hat{\phi}}} \leq \left(\sigma_{\hat{\phi}} \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)} + z_0\,\sigma_{\hat{\phi}}\right) \frac{1}{\sigma_{\hat{\phi}}}\right] =$$

$$= 1 - P_{\hat{F}}\left[\frac{\hat{\phi}^* - \hat{\phi} + z_0\,\sigma_{\hat{\phi}}}{\sigma_{\hat{\phi}}} \leq z_0 + \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\right] =$$

$$= 1 - \Phi\left[z_0 + \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\right].$$

Como,

$$1 - P_{\hat{F}}\left[\hat{\phi}^* \leq \hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\right] = 1 - \hat{H}\left[\hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\right],$$

quando $B \to +\infty$, vem,

$$1 - \hat{H}\left[\hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\right] = 1 - \Phi\left[z_0 + \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\right] \Leftrightarrow$$

$$\Leftrightarrow \hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)} = \hat{H}^{-1}\left\{\Phi\left[z_0 + \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\right]\right\}. \qquad (3.86)$$

Então, um intervalo de confiança "Bootstrap" para ϕ, a $(1-2\alpha)100\%$, é dado por,

$$\left[\hat{H}^{-1}\left\{\Phi\left[z_0 + \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\right]\right\} \;;\; \hat{H}^{-1}\left\{\Phi\left[z_0 + \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\right]\right\}\right]. \qquad (3.87)$$

Como a igualdade $\hat{H}(s) = P_{\hat{F}}(\hat{\phi}^* \leq s) \equiv P_*(\hat{\phi}^* \leq s)$ só é válida quando $B \to +\infty$, ou seja, como $\hat{H}(s)$ não é a verdadeira função de distribuição "Bootstrap" de $\hat{\phi}$, mas apenas uma sua estimativa, o nível de confiança do intervalo (3.87) é aproximado e não exacto.

Falta agora passar de um intervalo de confiança em ϕ para um intervalo de confiança em θ. De (3.33), tira-se $\theta = g^{-1}(\phi)$, logo, aplicando a função inversa g^{-1} aos limites do intervalo (3.87),

$$g^{-1}\left(\hat{H}^{-1}\left\{\Phi\left[z_0 + \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\right]\right\}\right) =$$

$$= g^{-1}\left(g\left\{\hat{G}^{-1}\left[\Phi\left(z_0 + \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\right)\right]\right\}\right) =$$

$$= \hat{G}^{-1}\left(\Phi\left[z_0 + \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\right]\right), \qquad (3.88)$$

e,

$$g^{-1}\left(\hat{H}^{-1}\left\{\Phi\left[z_0 + \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\right]\right\}\right) =$$

$$= g^{-1}\left(g\left\{\hat{G}^{-1}\left[\Phi\left(z_0 + \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\right)\right]\right\}\right) =$$

$$= \hat{G}^{-1}\left(\Phi\left[z_0 + \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\right]\right). \qquad (3.89)$$

Os resultados (3.88) e (3.89) assentam nas relações (3.46) e (3.47).

Então, de (3.88) e (3.89), pode sair (finalmente) o intervalo de confiança a $(1-2\alpha)100\%$ (aproximadamente) para θ, dado pelo método dos percentis corrigido do enviesamento e da aceleração da variância:

$$\left[\hat{\theta}_{BC_a}(LI)\,;\,\hat{\theta}_{BC_a}(LS)\right] =$$

$$= \left[\hat{G}^{-1}\left(\Phi\left[z_0 + \frac{z_0 + z_\alpha}{1 - a\,(z_0 + z_\alpha)}\right]\right)\,;\,\hat{G}^{-1}\left(\Phi\left[z_0 + \frac{z_0 + z_{1-\alpha}}{1 - a\,(z_0 + z_{1-\alpha})}\right]\right)\right]. \qquad (3.90)$$

Note-se que,

$$\hat{\theta}_{BC_a}(LI) = t_0 \ : \ t_0 = \inf_t \frac{\#\{\hat{\theta}^{*b} \leq t\}}{B} \geq \Phi\left[z_0 + \frac{z_0 + z_\alpha}{1 - a\ (z_0 + z_\alpha)}\right], \qquad (3.91)$$

e,

$$\hat{\theta}_{BC_a}(LS) = t_1 \ : \ t_1 = \inf_t \frac{\#\{\hat{\theta}^{*b} \leq t\}}{B} \geq \Phi\left[z_0 + \frac{z_0 + z_{1-\alpha}}{1 - a\ (z_0 + z_{1-\alpha})}\right]. \qquad (3.92)$$

Se o desvio padrão do estimador for constante, isto é, se a = 0, então $\hat{\theta}_{BC_a}(LI) = \hat{\theta}_{BC}(LI)$ e $\hat{\theta}_{BC_a}(LS) = \hat{\theta}_{BC}(LS)$, ou seja, o método dos percentis corrigido do enviesamento e da aceleração da variância conduz ao mesmo resultado que o método dos percentis corrigido do enviesamento.

Nas hipóteses (3.64), (3.65) e (3.66) que levaram à construção do intervalo (3.90), apresentaram-se duas constantes: a constante de enviesamento, z_0, e a constante de aceleração "a". Coloca-se agora o problema de determinar essas duas constantes, por forma a construir o intervalo de confiança (3.90).

Quanto à constante de enviesamento, o problema resolve-se facilmente com a expressão (3.53), deduzida no âmbito do método dos percentis corrigido do enviesamento e que, de forma similar, pode ser deduzida no contexto do método dos percentis corrigido do enviesamento e da aceleração da variância.

Quanto à constante de aceleração o problema torna-se mais difícil de resolver, sendo esta << ... *a parte mais delicada do estudo* >> como refere [Murteira (1988b) - pg. 12-13].

Efron começa por mostrar, no caso paramétrico unidimensional, que,

$$a = (1/6)\ SKEW_{\theta=\hat{\theta}}(\dot{L}_\theta), \qquad (3.93)$$

é uma boa aproximação para a constante de aceleração [veja-se Efron (1987) - pg. 174-176].

Na expressão (3.93), tem-se,

$$SKEW_{\theta=\hat{\theta}}(X) = \frac{E\{[X - E(X)]^3\}}{\left[E\{[X - E(X)]^2\}\right]^{3/2}}\Bigg|_{\theta=\hat{\theta}}, \qquad (3.94)$$

ou seja, é o coeficiente de assimetria de uma variável aleatória X, tomado em $\theta = \hat{\theta}$, e,

$$\dot{L}_\theta(\hat{\theta}) = \frac{\partial}{\partial\theta} \ln f_\theta(\hat{\theta}), \qquad (3.95)$$

onde $f_\theta(\hat{\theta})$ é a função de densidade de probabilidade do estimador $\hat{\theta}$.

Em seguida, Efron generaliza (3.93) para o caso multiparamétrico, apresentando um resultado para o caso concreto da família exponencial [veja-se Efron (1987) - pg. 176-177]. A aproximação para a constante de aceleração, no caso não paramétrico, deriva precisamente de uma adaptação deste resultado, provado para a família exponencial, aos domínios não paramétricos,

$$a = \frac{1}{6} \frac{\sum_{i=1}^{n} U_i^3}{\left[\sum_{i=1}^{n} U_i^2\right]^{3/2}}, \qquad (3.96)$$

onde,

$$U_i = (n - 1)\,[\hat{\theta}_\bullet - \hat{\theta}_{(i)}],\ i = 1, 2, \ldots, n, \qquad (3.97)$$

$$\hat{\theta}_{(i)} = \hat{\theta}(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n),\ i = 1, 2, \ldots, n, \qquad (3.98)$$

$$\hat{\theta}_{\cdot} = \sum_{i=1}^{n} (1/n)\ \hat{\theta}_{(i)} . \qquad (3.99)$$

As condições (3.97), (3.98) e (3.99), definem os U_i, i = 1, 2, ..., n, como uma versão da função de influência empírica, a qual mede a influência que a observação x_i tem no cálculo do estimador $\hat{\theta}$ [veja-se Murteira (1988b) - pg. 12-13].

Em termos teóricos, a função de influência empírica define-se por,

$$IF(x) = \lim_{\varepsilon \to 0} \frac{\theta[(1 - \delta)\ F + \varepsilon\ \delta(x)] - \theta(F)}{\varepsilon} , \qquad (3.100)$$

onde,

$$\delta(x) = \begin{cases} 1 , & X = x \\ 0 , & X \neq x \end{cases} .$$

Se, em (3.100), se substituir F por $\hat{F}$ e ε por $-1/(n - 1)$ (em vez de $\lim_{\varepsilon \to 0}$), obtem-se praticamente (3.97) [veja-se Murteira (1988) - pg. 13].

Se as expressões (3.64) e (3.65) forem exactas, então o intervalo de confiança dado pelo método dos percentis corrigido do enviesamento e da aceleração da variância coincide com o intervalo exacto (3.1) (esquecendo que B assume, na prática, um valor finito e que o valor da constante "a" é uma aproximação). Caso contrário, o intervalo (3.90) é aproximado e tem-se que as diferenças $|\hat{\theta}_{BC_a}(LI) - \theta_{LI}|$ e $|\hat{\theta}_{BC_a}(LS) - \theta_{LS}|$ são $O_p(n^{-3/2})$ (pelo menos, em domínios paramétricos unidimensionais, podendo conjecturar-se que também o sejam em domínios não paramétricos) [veja-se Dicicio e Romano (1988) - pg. 341-342], o que permite concluir por uma nítida melhoria em relação aos métodos dos percentis e dos percentis corrigido do enviesamento, cujas diferenças são $O_p(n^{-1})$, como foi explicitado nos pontos 3.2 e 3.3.2 do presente

estudo. Não se alargam mais as comparações entre os métodos "Bootstrap" já apresentados, dado que o ponto 3.6 é dedicado em exclusivo a essas comparações.

Finalmente, refira-se que, para a construção dos intervalos de confiança "Bootstrap" já apresentados (métodos dos percentis, dos percentis corrigido do enviesamento e dos percentis corrigido do enviesamento e da aceleração da variância), não é necessário conhecer a correcta transformação g - esta é apenas um instrumento auxiliar na dedução e justificação dos citados intervalos [veja-se Efron (1987) - pg. 172].

3.4 - O "BOOTSTRAP - t"

Se a distribuição do universo for normal, é conhecido o seguinte resultado,

$$\frac{\bar{X} - E(X)}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}, \qquad (3.101)$$

onde $\bar{X} = \sum_{i=1}^{n} (1/n)\ X_i$, é a média da amostra, $S^2 = \sum_{i=1}^{n} [1/(n-1)](X_i - \bar{X})^2$, é a variância (corrigida) da amostra e $t_{(n-1)}$ representa a distribuição t de Student, com n-1 graus de liberdade.

Em universos normais, a expressão (3.101) é vulgarmente utilizada para construir intervalos de confiança para a média do universo, E(X). Quando se passa de um universo normal para um universo desconhecido e o parâmetro a estimar $\theta \equiv \theta(F)$ não é, necessariamente, a média do universo, a distribuição de (3.101) passa a ser desconhecida, o que impossibilita o uso da expressão na construção de intervalos de confiança para o parâmetro θ.

O "Bootstrap" permite ultrapassar este impasse, ao construir uma expressão similar a (3.101), que pode ser utilizada em domínios não paramétricos para construir intervalos de confiança sobre o parâmetro θ.

Seja, então, a situação definida por (2.1), (2.2), (2.3) e (2.4), com a distribuição do universo, F, desconhecida (caso não paramétrico). O objectivo é

construir um intervalo de confiança a $(1-2\alpha)100\%$, $0 < \alpha < 0.5$, para o parâmetro $\theta \equiv \theta(F)$. Uma alternativa "Bootstrap" para constuir este intervalo baseia-se na seguinte estatística,

$$T = \frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}}, \tag{3.102}$$

onde $\hat{\sigma}_{\hat{\theta}}$ é um estimador consistente de σ (desvio padrão de $\hat{\theta}$).

Na verdade, sendo $t_\alpha : P(T \leq t_\alpha) = \alpha$ e $t_{1-\alpha} : P(T \leq t_{1-\alpha}) = 1-\alpha$, $0 < \alpha < 0.5$, tem-se,

$$P(t_\alpha \leq T \leq t_{1-\alpha}) = 1-2\alpha \Leftrightarrow$$

$$\Leftrightarrow P(t_\alpha \leq \frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}} \leq t_{1-\alpha}) = 1-2\alpha \Leftrightarrow$$

$$\Leftrightarrow P(-\hat{\theta} + t_\alpha \hat{\sigma}_{\hat{\theta}} \leq -\theta \leq -\hat{\theta} + t_{1-\alpha} \hat{\sigma}_{\hat{\theta}}) = 1-2\alpha \Leftrightarrow$$

$$\Leftrightarrow P(\hat{\theta} - t_{1-\alpha} \hat{\sigma}_{\hat{\theta}} \leq \theta \leq \hat{\theta} - t_\alpha \hat{\sigma}_{\hat{\theta}}) = 1-2\alpha .$$

Este resultado leva a apresentar,

$$[\hat{\theta} - t_{1-\alpha} \hat{\sigma}_{\hat{\theta}} \; ; \; \hat{\theta} - t_\alpha \hat{\sigma}_{\hat{\theta}}], \tag{3.103}$$

como intervalo de confiança a $(1-2\alpha)100\%$ para θ.

Como F é desconhecida, não é possível saber qual a exacta distribuição de T, logo não se consegue construir (3.103). No entanto, pode determinar-se a distribuição "Bootstrap" de T, isto é, a distribuição de,

$$T^* = \frac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}_{BOOT}} . \tag{3.104}$$

Para tal, recorre-se a um algoritmo de Monte Carlo:

1º) A partir da particular amostra observada $(x_1, x_2, ..., x_n)$, realizam-se B extracções (B suficientemente grande) com reposição, por forma a construir B amostras "Bootstrap" de dimensão n, $(x_1^{*b}, x_2^{*b}, ..., x_n^{*b})$, b = 1, 2, ..., B.

2º) Para cada amostra, calcula-se o valor "Bootstrap" de $T^{*b} = \dfrac{\hat{\theta}(x_1^{*b}, x_2^{*b}, ..., x_n^{*b}) - \hat{\theta}(x_1, x_2, ..., x_n)}{\hat{\sigma}_{BOOT}}$, b = 1, 2, ..., B.

Assim, fica-se com um conjunto de B valores "Bootstrap" de T: $T^{*1}, T^{*2}, ..., T^{*B}$.

3º) A função de distribuição empírica "Bootstrap" de T constrói-se facilmente,

$$\hat{D}(t) = \frac{\#\{T^{*b} \le t\}}{B}, \quad 1 \le b \le B, \quad -\infty < t < +\infty. \tag{3.105}$$

Note-se que a igualdade $\hat{D}(t) = P_{\hat{F}}(T^* \le t) \equiv P_*(T^* \le t)$ só é válida quando $B \to +\infty$, ou seja, $\hat{D}(t)$ não é a verdadeira função de distribuição "Bootstrap" de T, já que, não se conhece a exacta distribuição de (3.104) - esta sim, a verdadeira distribuição "Bootstrap" de T - mas apenas uma sua estimativa.

O resultado (3.105) vai permitir encontrar os valores "Bootstrap" para t_α e $t_{1-\alpha}$,

$$t_\alpha^* = t_0 \ : \ t_0 = \inf_t \frac{\#\{T^{*b} \le t\}}{B} \ge \alpha, \tag{3.106}$$

$$t_{1-\alpha}^* = t_1 \ : \ t_1 = \inf_t \frac{\#\{T^{*b} \le t\}}{B} \ge 1-\alpha. \tag{3.107}$$

Pode, então, construir-se a estimativa "Bootstrap" de (3.103),

$$\left[\hat{\theta}_{Bt}(LI)\ ;\ \hat{\theta}_{Bt}(LS)\right] = \left[\hat{\theta} - t^{*}_{1-\alpha}\,\hat{\sigma}_{BOOT}\ ;\ \hat{\theta} - t^{*}_{\alpha}\,\hat{\sigma}_{BOOT}\right]. \qquad (3.108)$$

A expressão (3.108) é o intervalo de confiança "Bootstrap - t" a $(1-2\alpha)100\%$ (aproximadamente, dado que $\hat{D}(t)$ não é exactamente a função de distribuição "Bootstrap" de T) para θ.

Os intervalos "Bootstrap - t", se bem que não tenham uma justificação teórica profunda, como os intervalos deduzidos por Efron (dos percentis, dos percentis corrigido do enviesamento e dos percentis corrigido do enviesamento e da aceleração da variância), o certo é que têm proporcionado bons resultados, possuindo algumas propriedades (as quais serão desenvolvidas no ponto 3.6) que os colocam na primeira linha dos intervalos de confiança "Bootstrap".

3.5 - O "BOOTSTRAP" ITERATIVO

Os métodos de construção de intervalos de confiança, deduzidos por Efron e o "Bootstrap - t", partem do princípio de que é possível encontrar uma variável pivô (ou seja, uma variável que depende de θ, mas cuja distribuição é independente de θ), a qual desempenha um papel primordial na construção dos intervalos de confiança:

- para o método dos percentis, têm-se, como variáveis pivô, $\frac{\hat{\phi} - \phi}{\tau} \sim N(0,1)$ e $\frac{\hat{\phi}^* - \hat{\phi}}{\tau} \underset{*}{\sim} N(0,1)$;

- para o método dos percentis corrigido do enviesamento, têm-se, $\frac{\hat{\phi} - (\phi - z_0\,\tau)}{\tau} \sim N(0,1)$ e $\frac{\hat{\phi}^* - (\hat{\phi} - z_0\,\tau)}{\tau} \underset{*}{\sim} N(0,1)$, como variáveis pivô;

- para o método dos percentis corrigido do enviesamento e da aceleração da variância, têm-se, $\frac{\hat{\phi} - (\phi - z_0\,\tau\,\sigma_\phi)}{\tau\,\sigma_\phi} \sim N(0,1)$ e $\frac{\hat{\phi}^* - (\hat{\phi} - z_0\,\tau\,\sigma_{\hat{\phi}})}{\tau\,\sigma_{\hat{\phi}}} \underset{*}{\sim} N(0,1)$, como variáveis pivô;

- para o "Bootstrap - t", têm-se, $\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}}$ e $\frac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}_{BOOT}}$, como variáveis pivô.

No entanto, na maioria dos casos, não há uma certeza absoluta sobre se as variáveis atrás referidas são mesmo variáveis pivô, podendo encontrar-se muito longe de o serem. O problema ganha uma evidência especial em

domínios não paramétricos, onde o elevado grau de desconhecimento leva a considerar esta ou aquela variável como pivô, unicamente por suposição e hipótese de partida. Se a variável considerada não for exacta ou aproximadamente pivô, então o nível de confiança real do intervalo de confiança construído pode ser substancialmente diferente do nível de confiança desejado $(1-2\alpha)100\%$, isto é, pode haver um significativo erro no nível de confiança.

Beran vem atacar este problema, propondo um novo método de construção de intervalos de confiança, baseado no "Bootstrap", o qual procura, através de sucessivas iterações, construir uma variável que se aproxime o mais possível de uma variável pivô [veja-se Beran (1987)]. Assim, torna-se possível construir um intervalo de confiança onde o erro no nível de significância não seja tão significativo.

Considere-se, então, uma situação definida por (2.1) e (2.2). A distribuição do universo, aqui representada pela respectiva função de distribuição, F, é desconhecida (está-se no caso não paramétrico). Pretende-se construir um intervalo de confiança a $(1-2\alpha)100\%$, $0 < \alpha < 0.5$ para o parâmetro $\theta \equiv \theta(F)$.

A variável aleatória utilizada na construção do intervalo de confiança é a seguinte,

$$R \equiv R_n(\theta) \equiv R_n(X_1, X_2, ..., X_n ; \theta). \qquad (3.109)$$

Seja Θ o conjunto de todos os possíveis valores do parâmetro θ e designe-se pela letra u o valor corrente do parâmetro θ.

Se $R_n(\theta)$ for uma variável pivô, com distribuição conhecida (por mera hipótese), o intervalo de confiança a $(1-2\alpha)100\%$ para θ é dado por,

$$\{u \in \Theta: \alpha < J_n[R_n(u)] < 1 - \alpha\} =$$

$$= \{u \in \Theta: J_n^{-1}(\alpha) < R_n(u) < J_n^{-1}(1-\alpha)\}, \qquad (3.110)$$

onde $J_n \equiv J_n(\cdot,F)$ é a função de distribuição, contínua à esquerda, de $R_n(\theta)$ e $J_n^{-1}(\alpha)$ e $J_n^{-1}(1-\alpha)$, os quantis α e $1-\alpha$, respectivamente, da distribuição de $R_n(\theta)$[7]. Repare-se que a função de distribuição $J_n(\cdot,F)$ só é conhecida no caso em que se conhece F.

Na maioria dos casos, a variável aleatória $R_n(\theta)$ não é uma variável pivô, mas, apesar de não se conhecer a sua distribuição (desconhece-se $J_n(\cdot,F)$), a mesma pode ser estimada consistentemente, a partir da amostra (neste caso, diz-se que $R_n(\theta)$ é a raiz do intervalo de confiança).

O "Bootstrap" permite ultrapassar esta situação ao estimar a distribuição da raiz $R_n(\theta)$.

Seja, então,

$$\hat{J}_n \equiv J_n(\cdot,\hat{F}), \qquad (3.111)$$

a função de distribuição "Bootstrap" da raiz $R_n(\theta)$. Um intervalo de confiança a $(1-2\alpha)100\%$ (aproximadamente) para θ é dado por,

$$B_n = \{u \in \Theta: \alpha \le \hat{J}_n[R_n(u)] \le 1 - \alpha\} =$$

$$= \{u \in \Theta: \hat{J}_n^{-1}(\alpha) \le R_n(u) \le \hat{J}_n^{-1}(1 - \alpha)\}. \qquad (3.112)$$

---

[7] Quando se está a construir estes intervalos de confiança, $R_n(u)$ varia apenas em função de $u \in \Theta$, estando fixa a amostra observada $(x_1, x_2, \ldots, x_n)$, isto é, $R_n(u) = R_n(x_1, x_2, \ldots, x_n; u)$. O mesmo se torna válido para os intervalos de confiança que a seguir são deduzidos. O emprego de $R_n(u)$ em vez de $R_n(x_1, x_2, \ldots, x_n ; u)$ destina-se apenas a não tornar a notação ainda mais "pesada", tendo-se sempre subjacente que o vector aleatório $(X_1, X_2, \ldots, X_n)$ está fixo em $(x_1, x_2, \ldots, x_n)$.

Repare-se que o nível de confiança de (3.112) não é exactamente $(1-2\alpha)100\%$, porque a raiz $R_n(\theta)$ não é uma variável pivô. No entanto, o nível de confiança assintótico (quando $n \to +\infty$) de (3.112) é mesmo $(1-2\alpha)100\%$, se for verificada a seguinte condição [veja-se Beran (1987) - pg. 459]: seja $\{F_n \in \mathfrak{F}\}$[8] uma qualquer sucessão em que F_n converge para F, numa métrica d, então $J_n(\cdot,F_n)$ converge fracamente para a função de distribuição contínua $J(\cdot,F)$, a qual depende apenas de F e não da sucessão $\{F_n\}$. Esta condição e a consistência da distribuição empírica da amostra $\hat{F}$, como estimador de F, implicam que,

$$\sup_x \| J_n(x,\hat{F}) - J(x,F) \| \xrightarrow{P} 0.$$

Como a condição atrás enunciada também assegura a convergência de $J_n \equiv J_n(\cdot,F)$ para $J(\cdot,F)$, tem-se que a distribuição de $\hat{J}_n[R_n(\theta)]$ converge fracamente para uma distribuição uniforme no intervalo [0 ; 1], o que, tendo em atenção (3.112), permite afirmar que este intervalo tem um nível de confiança assintótico $(1-2\alpha)100\%$ [veja-se Beran (1987) - pg. 459].

Note-se que, se $R_n(X_1, X_2, ..., X_n ; \theta) = \hat{\theta}(X_1, X_2, ..., X_n)$, o intervalo (3.112) não é mais do que o intervalo obtido pelo método dos percentis,

$$B_n = \{u \in \Theta: \alpha \le \hat{G}(u) \le 1 - \alpha\} =$$

$$= \{u \in \Theta: \hat{G}^{-1}(\alpha) \le u \le \hat{G}^{-1}(1 - \alpha)\} =$$

$$= \left[\hat{\theta}_{PER}(LI) ; \hat{\theta}_{PER}(LS)\right].$$

O erro no nível de confiança do intervalo (3.112), ou seja, a diferença entre o nível de confiança real de (3.112) e o nível de confiança desejado

[8] $\mathfrak{F}$ é a família de distribuições a que pertence a "verdadeira" distribuição do universo, F.

$(1-2\alpha)100\%$, depende: da raiz escolhida, da distribuição da amostra e do método utilizado para estimar $J_n(\cdot,F)$ (no caso presente, o "Bootstrap").

O erro no nível de confiança pode ser considerável, mesmo em amostras de dimensão relativamente grande. Para diminuir este erro, os intervalos de confiança devem gerar-se, não a partir da raiz inicial, mas sim, de uma outra raiz que resulta de sucessivas transformações da raiz inicial.

No intervalo (3.112), faça-se,

$$R_{n,1}(\theta) = \hat{J}_n[R_n(\theta)], \tag{3.113}$$

ou seja, passa-se da raiz $R_n(\theta)$ para a raiz $R_{n,1}(\theta)$, a qual irá servir para construir um novo intervalo de confiança.

A passagem de $R_n(\theta)$ para $R_{n,1}(\theta)$ pode justificar-se da seguinte forma: é plausível que a distribuição de $R_{n,1}(\theta)$ dependa, de uma maneira menos forte, de F, do que a distribuição de $R_n(\theta)$ [veja-se Beran (1987) - pg. 459]. Como as distribuições de $R_n(\theta)$ e $R_{n,1}(\theta)$ dependem de F através de $X_1, X_2, \ldots, X_n$ e de $\theta \equiv \theta(F)$, então dizer que $R_{n,1}(\theta)$ depende, de uma maneira menos forte, de F, do que $R_n(\theta)$, significa que $R_{n,1}(\theta)$ também depende, de uma maneira menos forte, de $\theta \equiv \theta(F)$, do que $R_n(\theta)$. Por outras palavras, é plausível que $R_{n,1}(\theta)$ se aproxime mais de uma variável pivô do que $R_n(\theta)$.

A passagem de $R_n(\theta)$ para $R_{n,1}(\theta)$, através da função de distribuição estimada de $R_n(\theta)$ (neste caso, a função de distribuição "Bootstrap" de $R_n(\theta)$), chama-se "prepivoting" [veja-se Beran (1987) - pg. 459] e não é mais do que a tentativa de transformar a raiz inicial noutra raiz, que se aproxime mais de uma variável pivô, permitindo construir um intervalo de confiança com menor erro no nível de confiança.

Para construir um novo intervalo de confiança, a partir de $R_{n,1}(\theta)$, seja $J_{n,1} \equiv J_{n,1}(\cdot,F)$ a função de distribuição de $R_{n,1}(\theta)$ e $\hat{J}_{n,1} \equiv J_{n,1}(\cdot,\hat{F})$ a correspondente estimativa "Bootstrap". Nestas condições, um novo intervalo de confiança para θ é,

$$B_{n,1} = \{u \in \Theta: \alpha \le \hat{J}_{n,1}[R_{n,1}(u)] \le 1 - \alpha\} =$$

$$= \{u \in \Theta: \alpha \le \hat{J}_{n,1}(\hat{J}_n[R_n(u)]) \le 1 - \alpha\} =$$

$$= \left\{u \in \Theta: \hat{J}_n^{-1}\left[\hat{J}_{n,1}^{-1}(\alpha)\right] \le R_n(u) \le \hat{J}_n^{-1}\left[\hat{J}_{n,1}^{-1}(1-\alpha)\right]\right\}. \qquad (3.114)$$

É plausível que o nível de confiança real de (3.114) esteja mais perto de $(1-2\alpha)100\%$, do que o nível de confiança real de (3.112).

A operação de "prepivoting" pode ser repetida iterativamente, dando origem aos seguintes intervalos de confiança,

$$B_{n,j} = \{u \in \Theta: \alpha \le \hat{J}_{n,j}[R_{n,j}(u)] \le 1 - \alpha\}, \; j = 1, 2, \ldots, \qquad (3.115)$$

onde,

$$R_{n,j+1}(\theta) \equiv \hat{J}_{n,j}[R_{n,j}(\theta)], \; j = 1, 2, \ldots, \qquad (3.116)$$

$$\hat{J}_{n,j} \equiv J_{n,j}(\cdot,\hat{F}), \; j = 1, 2, \ldots, \qquad (3.117)$$

$J_{n,j} \equiv J_{n,j}(\cdot,F)$ - função de distribuição da raiz $R_{n,j}(\theta)$, $j = 1, 2, \ldots$. (3.118)

Nos casos regulares, o erro no nível de confiança dos $B_{n,j}$ diminui com o aumento de j, ou seja, diminui com as sucessivas operações de "prepivoting" [veja-se Beran (1987) - pg. 460].

Até agora, os intervalos $B_{n,j}$ foram deduzidos supondo que as distribuições "Bootstrap" $\hat{J}_{n,j}$ são perfeitamente conhecidas. Ora, tal não acontece em domínios não paramétricos, tendo de se recorrer ao já conhecido

algoritmo de Monte Carlo. Vai exemplificar-se a aplicação deste algoritmo, quando se pretende construir o intervalo de confiança $B_{n,1}$, caso em que se necessita do conhecimento das funções de distribuição "Bootstrap" $\hat{J}_n$ e $\hat{J}_{n,1}$. Defina-se, então,

$\underline{X}_n \equiv (X_1, X_2, ..., X_n)$ - amostra aleatória inicial gerada a partir da distribuição F.

$\underline{X}_n^* \equiv (X_1^*, X_2^*, ..., X_n^*)$ - amostra "Bootstrap" genérica, gerada a partir de tiragens com reposição da particular amostra observada, $\underline{x}_n \equiv (x_1, x_2, ..., x_n)$, ou seja, gerada a partir de $\hat{F}$.

Note-se que os X_j^*, $j = 1, 2, ..., n$, são condicionalmente independentes, dada a amostra inicial $\underline{x}_n$.

$\underline{X}_n^{**} \equiv (X_1^{**}, X_2^{**}, ..., X_n^{**})$ - amostra "Bootstrap" gerada a partir de tiragens com reposição da particular amostra, $\underline{x}_n^* \equiv (x_1^*, x_2^*, ..., x_n^*)$, ou seja, gerada a partir de $\hat{F}^*$, onde $\hat{F}^*$ é a função de distribuição empírica da amostra "Bootstrap" $\underline{x}_n^*$.

Note-se que os X_j^{**}, $j = 1, 2, ..., n$, são condicionalmente independentes, dadas a amostra inicial $\underline{x}_n$ e a amostra "Bootstrap" $\underline{x}_n^*$.

$\hat{\theta}_n \equiv \hat{\theta} \equiv \hat{\theta}(x_1, x_2, ..., x_n)$ - estimativa do parâmetro θ, baseada na amostra inicial $\underline{x}_n$.

$\hat{\theta}_n^* \equiv \hat{\theta}^* \equiv \hat{\theta}(x_1^*, x_2^*, ..., x_n^*)$ - estimativa do parâmetro θ, baseada na amostra "Bootstrap" $\underline{x}_n^*$.

De acordo com estas definições, tem-se,

$$J_n \equiv J_n(x,F) = P[R_n(\underline{X}_n,\theta) < x \mid F], \tag{3.119}$$

e,

$$J_{n,1} \equiv J_{n,1}(x,F) = P[R_{n,1}(\underline{X}_n,\theta) < x \mid F] =$$

$$= P\{\hat{J}_n[R_n(\underline{X}_n,\theta)] < x \mid F\} =$$

$$= P\{P[R_n(\underline{X}_n^*,\hat{\theta}_n) < R_n(\underline{X}_n,\theta) \mid \hat{F}] < x \mid F\}. \tag{3.120}$$

Note-se que $R_n(\underline{X}_n^*,\hat{\theta}_n)$ é a aproximação "Bootstrap" de $R_n(\underline{X}_n,\theta)$.

Como as funções de distribuição J_n e $J_{n,1}$ são desconhecidas, importa passar para as suas estimativas "Bootstrap",

$$\hat{J}_n \equiv J_n(x,\hat{F}) = P[R_n(\underline{X}_n^*,\hat{\theta}_n) < x \mid \hat{F}], \tag{3.121}$$

$$\hat{J}_{n,1} \equiv J_{n,1}(x,\hat{F}) = P\{P[R_n(\underline{X}_n^{**},\hat{\theta}_n^*) < R_n(\underline{X}_n^*,\hat{\theta}_n) \mid \hat{F}^*] < x \mid \hat{F}\}. \tag{3.122}$$

Repare-se que $R_n(\underline{X}_n^{**},\hat{\theta}_n^*)$ é a aproximação "Bootstrap" de $R_n(\underline{X}_n^*,\hat{\theta}_n)$.

Chegado a este ponto, o problema reside no facto de, em universos não paramétricos, não se conhecerem exactamente as funções de distribuição "Bootstrap" $\hat{J}_n$ e $\hat{J}_{n,1}$. A solução está em estimá-las, de acordo com os seguintes algoritmos de Monte Carlo:

i) Estimação de $\hat{J}_n$.

1º) A partir da amostra observada $\underline{x}_n$, realizam-se extracções com reposição, por forma a gerar B_1 (B_1 suficientemente grande) amostras "Bootstrap" de dimensão n: $\underline{x}_n^{*1}, \underline{x}_n^{*2}, ..., \underline{x}_n^{*B_1}$.

Note-se que as B_1 amostras "Bootstrap" são condicionalmente independentes, dada a amostra inicial $\underline{x}_n$.

2º) Para cada amostra "Bootstrap", $\underline{x}_n^{*b}$, b = 1, 2, ..., B_1, calcula-se $R_n(\underline{x}_n^{*b}, \hat{\theta}_n)$, b = 1, 2, ..., B_1.

A distribuição empírica dos valores $\{R_n(\underline{x}_n^{*b}, \hat{\theta}_n) : 1 \leq b \leq B_1\}$ aproxima $\hat{J}_n$, para valores suficientemente grandes de B_1.

ii) Estimação de $\hat{J}_{n,1}$.

1º) A partir de cada uma das amostras "Bootstrap", $\underline{x}_n^{*b}$, b = 1, 2, ..., B_1, realizam-se extracções com reposição, por forma a gerar B_2 amostras "Bootstrap", de dimensão n, de "segunda ordem": $\underline{x}_{n,b}^{**1}$, $\underline{x}_{n,b}^{**2}$, ..., $\underline{x}_{n,b}^{**B_2}$, b = 1, 2, ..., B_1.

Note-se que as B_1B_2 amostras "Bootstrap" de "segunda ordem" são condicionalmente independentes, dadas a amostra inicial $\underline{x}_n$ e as amostras "Bootstrap", $\underline{x}_n^{*b}$, b = 1, 2, ..., B_1.

2º) Para cada amostra "Bootstrap", de "primeira ordem", $\underline{x}_n^{*b}$, b = 1, 2, ..., B_1, calcula-se $\hat{\theta}_n^{*b} \equiv \hat{\theta}(\underline{x}_n^{*b})$, b = 1, 2, ..., B_1.

3º) Para cada amostra "Bootstrap" de "segunda ordem", $\underline{x}_{n,b}^{**t}$, t = 1, 2, ..., B_2, b = 1, 2, ..., B_1, calcula-se $R_n(\underline{x}_{n,b}^{**t}, \hat{\theta}_n^{*b})$, t = 1, 2, ..., B_2, b = 1, 2, ..., B_1.

4º) Faz-se,

$$Z_b = \frac{\#\{R_n(\underline{x}^{**t}_{n,b}, \hat{\theta}^{*b}_n) \leq R_n(\underline{x}^{*b}_n, \hat{\theta}_n) \ : \ 1 \leq t \leq B_2\}}{B_2}, \ b = 1, 2, ..., B_1.$$

A distribuição empírica dos valores $\{Z_b, b = 1, 2, ..., B_1\}$ aproxima $\hat{J}_{n,1}$, para valores suficientemente grandes de B_1 e B_2. Este facto percebe-se melhor, se, na expressão de Z_b, se aplicar $\hat{J}_n$ a $R_n(\cdot,\cdot)$, ficando, assim, com,

$$Z_b = \frac{\#\{R_{n,1}(\underline{x}^{**t}_{n,b}, \hat{\theta}^{*b}_n) \leq R_{n,1}(\underline{x}^{*b}_n, \hat{\theta}_n) \ : \ 1 \leq t \leq B_2\}}{B_2}, \ b = 1, 2, ..., B_1.$$

Recorrendo a este algoritmo, os limites inferior e superior do intervalo (3.114) podem-se estimar da seguinte forma,

$$\hat{\theta}_{IT}(LI) = \hat{J}_n^{-1}\left[\hat{J}_{n,1}^{-1}(\alpha)\right] =$$

$$= y_0 \ : \ y_0 = \inf_y \frac{\#\{R_n(\underline{x}^{*b}_n, \hat{\theta}_n) \leq y\}}{B_1} \geq \left[\inf_x \frac{\#\{Z_b \leq x\}}{B_1} \geq \alpha\right], \qquad (3.123)$$

e,

$$\hat{\theta}_{IT}(LS) = \hat{J}_n^{-1}\left[\hat{J}_{n,1}^{-1}(1-\alpha)\right] =$$

$$= y_1 \ : \ y_1 = \inf_y \frac{\#\{R_n(\underline{x}^{*b}_n, \hat{\theta}_n) \leq y\}}{B_1} \geq \left[\inf_x \frac{\#\{Z_b \leq x\}}{B_1} \geq 1 - \alpha\right]. \qquad (3.124)$$

Como já se disse, quanto mais iterações de "prepivoting" se fizerem, menor será o erro no nível de confiança do respectivo intervalo $B_{n,j}$. No entanto, a tendência natural, que seria fazer $j \to +\infty$, esbarra com problemas de ordem computacional. Por exemplo, para o caso dos intervalos $B_{n,1}$, e fazendo

$B_1 = B_2 = 1000$, como aconselha Beran [veja-se Beran (1987) - pg. 461][9], o número de amostras envolvidas na estimação de $\hat{J}_{n,1}$ atinge um milhão, sendo este valor multiplicado por mil, por cada iteração a mais de "prepivoting" (pressupondo que se mantem em 1000 o número de novas amostras "Bootstrap" a gerar, a partir de cada amostra "Bootstrap" existente na fase anterior). Este acentuado crescimento dos cálculos a efectuar levanta problemas de disponibilidade de recursos e conduz, em termos práticos, a não se passar dos intervalos $B_{n,1}$.

Os problemas atrás focados têm motivado a procura de outras vias para a construção de intervalos de confiança, em domínios onde a informação relevante disponível é escassa. Uma dessas vias, bastante em foco nos últimos tempos, baseia-se nas conhecidas expansões de Edgeworth.

Não é pretensão deste estudo explanar as expansões de Edgeworth, nem, tão-pouco, apresentar os intervalos de confiança deduzidos por esta corrente teórica. Pretende, apenas, mostrar-se que, antes de serem dois métodos concorrentes, o "Bootstrap" e as expansões de Edgeworth têm, entre si, múltiplos pontos de contacto, podendo complementar-se na obtenção de melhores intervalos de confiança.

---

[9] Na construção de qualquer tipo de intervalo de confiança "Bootstrap", o número de amostras "Bootstrap" a construir deverá ser 1000, no mínimo, como recomenda Efron: « *A large number of bootstrap replications, B = 1000, in this case, is necessary to get reasonable accuracy in the tails of the distribution.* » [Efron (1982a) - pg. 78].
Posteriormente, Hall analisou, com maior profundidade, esta questão do número de amostras "Bootstrap" que é necessário gerar para construir intervalos de confiança e chegou à conclusão de que, se B (número de amostras "Bootstrap") não puder ser muito grande, as penalizações que daí advêm não se reflectem muito no erro no nível de confiança, mas antes, no comprimento do intervalo que tem tendência em aumentar. No entanto, estas conclusões não são gerais e resultam apenas da observação de alguns exemplos para o caso do "Bootstrap - t" [veja-se Hall (1986b) - pg. 1454].

Considere-se uma situação definida por (2.1) e (2.2). A distribuição do universo é desconhecida e o objectivo é construir um intervalo intervalo de confiança a $(1-2\alpha)100\%$, $0 < \alpha < 0.5$, para o parâmetro $\theta \equiv \theta(F)$.

A raiz do intervalo de confiança [veja-se (3.109)] é dada por,

$$R_n(\theta) \equiv R_n(X_1, X_2, \ldots, X_n ; \theta) = \sqrt{n}\,(\hat{\theta} - \theta). \qquad (3.125)$$

Suponha-se que a distribuição da raiz (3.125) é assintoticamente normal, de média nula e desvio padrão $SD \equiv SD(F)$.

Um estimador consistente de SD(F) é, naturalmente, o estimador "Bootstrap", $\hat{SD} \equiv SD(\hat{F})$.

Suponha-se, ainda, que é válida a seguinte expansão de Edgeworth, uniforme em x,

$$P\left\{\frac{R_n(\theta)}{SD(F)} < x\right\} = \Phi(x) + \sum_{j=1}^{2} n^{-(1/2)j}\, t_j(x,F) + o(n^{-1}), \qquad (3.126)$$

onde t_1 é uma função par de x e t_2 é uma função ímpar de x, as duas dependendo "suavemente" de F.

Nestas condições, é possível construir aproximações analíticas para a função de distribuição da raiz $R_n(\theta)$, $J_n \equiv J_n(\cdot,F)$, e para a função de distribuição da raiz "prepivoted" $R_{n,1}(\theta)$, $J_{n,1} \equiv J_{n,1}(\cdot,F)$. A título de exemplo, apresenta-se a aproximação analítica de $J_n(\cdot,F)$,

$$J_{EDG} = J_n(\cdot,F) = \Phi\left[\frac{x}{SD(F)}\right] + \sum_{j=1}^{2} n^{-(1/2)j}\, t_j\left(\frac{x}{SD(F)}, F\right) + o(n^{-1}),$$

cuja estimativa "Bootstrap" se obtem, substituindo F por $\hat{F}$,

$$\hat{J}_{EDG} = J_n(\cdot,\hat{F}) = \Phi\left[\frac{x}{SD(\hat{F})}\right] + \sum_{j=1}^{2} n^{-(1/2)j}\, t_j\left(\frac{x}{SD(\hat{F})}, \hat{F}\right) + o(n^{-1}),$$

o que exemplifica a complementaridade entre as expansões de Edgeworth e o "Bootstrap" [veja-se Beran (1987) - pg. 464-465].

As aproximações analíticas de $J_n(\cdot,F)$ e $J_{n,1}(\cdot,F)$ abrem a possibilidade de construir novos intervalos de confiança onde, em vez das estimativas "Bootstrap" directas das funções de distribuição $J_n(\cdot,F)$ e $J_{n,1}(\cdot,F)$, passam a ter-se as aproximações analíticas destas mesmas funções de distribuição, deduzidas com recurso às expansões de Edgeworth. Nesta linha de ideias, e tendo por base os intervalos $B_{n,1}$ [veja-se (3.114)], pode concluir-se [veja-se Beran (1987) - pg. 465] que:

- substituindo as estimativas "Bootstrap" $\hat{J}_n$ e $\hat{J}_{n,1}$, presentes em (3.114), pelas suas aproximações analíticas, deduzidas com base nas expansões de Edgeworth, se chega aos métodos propostos por Withers [veja-se Withers (1983) e Withers (1984)] e Hall [veja-se Hall (1983)];

- substituindo apenas a estimativa "Bootstrap" $\hat{J}_n$, presente em (3.114), por uma sua aproximação analítica, deduzida com base nas expansões de Edgeworth, e mantendo a estimativa "Bootstrap" $\hat{J}_{n,1}$, se cai no método proposto por Abramovitch e Singh [veja-se Abramovitch e Singh (1985)];

- substituindo apenas a estimativa "Bootstrap" $\hat{J}_{n,1}^{-1}$, presente em (3.114), por uma sua aproximação analítica, deduzida com base nas expansões inversas de Edgeworth, e mantendo a estimativa "Bootstrap" $\hat{J}_n$, se vai ter (aproximadamente) ao método proposto por Hall [veja-se Hall (1986a)].

Como pode ver-se, através desta exposição sumária, existe uma multiplicidade de hipóteses para construir intervalos de confiança, através da conjugação entre o "Bootstrap" e as expansões de Edgeworth, o que só demonstra a fecundidade destas duas linhas de investigação e o interesse de uma interligação entre ambas.

3.6 - COMPARAÇÕES ENTRE OS MÉTODOS "BOOTSTRAP" APRESENTADOS

Os cinco métodos "Bootstrap" apresentados, conduzem, geralmente, a diferentes intervalos de confiança. Deste modo, importa reter as propriedades de cada tipo de intervalos, para fazer as necessárias comparações e optar pelo método que proporcionar melhores "performances".

A análise comparativa dos diferentes intervalos de confiança "Bootstrap" depara com grandes obstáculos, já que, o "Bootstrap" é matéria recente, com muitos problemas ainda em aberto, em que se torna difícil deduzir resultados válidos sob condições muito gerais.

No caso concreto dos intervalos de confiança, a maioria das propriedades deduzidas têm-no sido para o caso paramétrico, extrapolando-se hipoteticamente a sua validade para o caso não paramétrico. Estas limitações condicionam a ligeira abordagem a que vai proceder-se.

Os limites inferior e superior do intervalo de confiança "verdadeiro" a $(1-2\alpha)100\%$ para θ (só possível de construir se se conhecer a distribuição F), dado pela expressão (3.1), são da forma,

$$\hat{\theta} + \hat{\sigma}\left[z^{(\alpha)} + \frac{A_n^{(\alpha)}}{n^{1/2}} + \frac{B_n^{(\alpha)}}{n} + \frac{C_n^{(\alpha)}}{n^{3/2}} + \cdots\right], \qquad (3.127)$$

[veja-se Efron (1987) - pg. 171], onde $\hat{\sigma}$ é uma estimativa do desvio padrão de $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ (muitas das vezes, $\hat{\sigma}$ é a estimativa baseada na quantidade

de informação de Fisher, $\hat{\sigma}_{IF}$, tal como foi apresentada no ponto 3.2), $A_n^{(\alpha)}$, $B_n^{(\alpha)}$, $C_n^{(\alpha)}$, ..., são constantes que dependem de n e α (os quais se supõem fixos) e $z^{(\alpha)}$ é um quantil da distribuição normal estandardizada dependente de α.

Uma forma de aquilatar a qualidade de cada intervalo "Bootstrap" reside em analisar a proximidade entre os seus limites inferior e superior e os limites inferior e superior do "verdadeiro" ou exacto intervalo de confiança. Quanto maior for essa proximidade, melhor será, evidentemente, a qualidade do intervalo "Bootstrap" em causa.

O intervalo de confiança tradicional ou "standard", dado pela expressão (3.9), é correcto em primeira ordem, no sentido em que o termo $\hat{\theta} + \hat{\sigma}\, z^{(\alpha)}$ domina assintoticamente a expressão (3.127) [veja-se Efron (1987) - pg. 171]. Nesta linha de ideias, as diferenças entre os limites inferiores e superiores dos intervalos "standard" e dos intervalos exactos, ou seja, $|\hat{\theta}_{TRA}(LI) - \theta_{LI}|$ e $|\hat{\theta}_{TRA}(LS) - \theta_{LS}|$, são $O_P(n^{-1})$. Exemplifique-se, para o caso do limite inferior.

Sabendo-se que θ_{LI} tem a forma da expressão (3.127) e que $\hat{\theta}_{TRA}(LI)$ "apanha" os termos $\hat{\theta} + \hat{\sigma}\, z^{(\alpha)}$, então,

$$|\hat{\theta}_{TRA}(LI) - \theta_{LI}| = \left|-\hat{\sigma}\left(\frac{A_n^{(\alpha)}}{n^{1/2}} + \frac{B_n^{(\alpha)}}{n} + \frac{C_n^{(\alpha)}}{n^{3/2}} + \ldots\right)\right|. \qquad (3.128)$$

Como $\hat{\sigma} = \hat{\sigma}_{IF}$ é da ordem $O_P(n^{-1/2})$ [veja-se Efron (1987) - pg. 176], então a expressão (3.128) é da ordem $O_P(n^{-1})$.

Os intervalos de confiança baseados nos métodos dos percentis e dos percentis corrigido do enviesamento não vêm trazer grandes melhorias, neste âmbito, em relação ao método tradicional - eles continuam a ter limites inferior e superior correctos apenas em primeira ordem [veja-se Dicicio e Romano (1988) - pg. 340 e Hall (1988b) - pg. 944], nos casos em que as

expressões (3.36) e (3.37), por parte do método dos percentis e (3.48) e (3.49), por parte do método dos percentis corrigido do enviesamento, não se verificam exactamente, mas apenas aproximadamente.

Os intervalos de confiança baseados no método dos percentis corrigido do enviesamento e da aceleração da variância (intervalos BC_a) contribuem para uma maior aproximação em relação aos extremos do intervalo exacto, alcançando a correcção de segunda ordem, no caso em que as expressões (3.64) e (3.65) não se verificam exactamente, mas apenas aproximadamente: « *In general z_0 and a do not equal zero and... make adjustments to the percentile method that are necessary to achieve second-order correctness* » [Efron (1987) - pg. 173].

A correcção em segunda ordem dos extremos dos intervalos BC_a , no sentido em que o termo $\hat{\theta} + \hat{\sigma}\left[z^{(\alpha)} + \frac{A_n^{(\alpha)}}{n^{1/2}}\right]$ domina assintoticamente a expressão (3.127), leva a que as diferenças $\left|\hat{\theta}_{BC_a}(LI) - \theta_{LI}\right|$ e $\left|\hat{\theta}_{BC_a}(LS) - \theta_{LS}\right|$ sejam $O_P(n^{-3/2})$ [veja-se Dicicio e Romano (1988) - pg. 341-342].

É certo que a correcção em segunda ordem dos extremos dos intervalos BC_a foi demonstrada, por Efron, apenas para o caso paramétrico unidimensional - « ... *for simple one-parameter problems, the BC_a intervals coincide through second order with the exact intervals* » [Efron (1987) - pg. 176] - conjecturando-se a sua validade para os domínios não paramétricos. Posteriormente, Hall generalizou a correcção em segunda ordem dos extremos dos intervalos BC_a ao caso multiparamétrico, quando a distribuição do universo pertence à família exponencial multivariada, e ao caso não paramétrico, quando $\hat{\theta}$ pode ser expresso como função de um vector de médias [veja-se Hall (1988b) - pg. 928].

Hall, que deduz praticamente os mesmos intervalos de confiança apresentados por Efron, embora recorrendo a outra forma de raciocínio que não passa pela teoria da transformação [veja-se Hall (1988b) - pg. 927-940], tende a atribuir grande importância aos intervalos de confiança baseados no "Bootstrap - t" (intervalos studendizados), os quais também possuem extremos correctos em segunda ordem [veja-se Hall (1988b) - pg. 944]. Chega mesmo a considerar os intervalos studentizados melhores do que os intervalos BC_a, na medida em que os primeiros permitem "apanhar" melhor os termos de terceira ordem, desde que a estimativa do desvio padrão, $\hat{\sigma}_{\hat{\theta}}$, tenha sido correctamente escolhida[10] [veja-se Hall (1988b) - pg. 929]. A importância dos termos de terceira ordem ganha relevância, porquanto, em intervalos bilaterais, o comprimento do intervalo é sobretudo influenciado por condições de terceira ordem; as condições de segunda ordem influenciam predominantemente a probabilidade de cobertura, ou seja, o nível de confiança real do intervalo [veja-se Hall (1988b) - pg. 929].

Numa "discussion" sobre o artigo de Hall que tem vindo a ser citado, Efron mostra algumas reticências em relação ao entusiasmo pelos intervalos studentizados, pondo em evidência algumas das suas deficiências que o fazem optar pelos intervalos BC_a: « *My original enthusiasm for bootstrap - t intervals, as naively expressed in Remark F of EFRON (1979) and slightly less naively in Section 10.10 of EFRON (1982), fondered on a list of their substancial drawbacks: noninvariance under transformations, occasional*

---

[10] Em (3.104), a estimativa escolhida para $\hat{\sigma}_{\hat{\theta}}$ foi $\hat{\sigma}_{BOOT}$, escolha esta que não é pacífica, não existindo, ainda, uma posição de consenso sobre qual a "melhor" estimativa de $\hat{\sigma}_{\hat{\theta}}$ a utilizar na construção de T^*.
Refira-se, no entanto, que, em domínios não paramétricos, a escolha de uma estimativa para $\hat{\sigma}_{\hat{\theta}}$ fica algo limitada, não se podendo, em princípio, apresentar algo muito melhor do que $\hat{\sigma}_{BOOT}$.

numerical instability and, worst of all, the need to compute auxiliary estimates of standard deviation $\hat{\sigma}$ *and* $\hat{\sigma}^*$. » [Hall (1988b) - "Discussion" de Efron - pg. 969].

Os intervalos de confiança baseados no "Bootstrap" iterativo (intervalos iterativos) são construídos com uma preocupação diferente da que presidiu à construção dos intervalos apresentados por Efron. Quanto a estes, procura-se que os respectivos limites inferior e superior se aproximem o mais possível dos limites inferior e superior do intervalo exacto, ou seja, procura-se a maior correcção possível na aproximação de (3.127) - daí a ênfase dada à correcção de segunda ordem dos limites dos intervalos BC_a. Quanto aos intervalos iterativos, a sua preocupação centra-se mais no erro cometido no nível de confiança, o qual pode ser significativo, a partir do momento em que a variável aleatória utilizada para construir os intervalos de confiança não é (nem aproximadamente) uma variável pivô - os intervalos deduzidos por Efron ladeiam este problema, na medida em que as suas hipóteses permitem a obtenção (exacta ou aproximada) de variáveis pivô normais, mas Beran, levantando grandes dúvidas sobre a plausibilidade dessas hipóteses, quando aplicadas à resolução de uma variada gama de problemas concretos, propõe um "prepivoting" iterativo que permita passar da variável aleatória inicial (raiz do intervalo de confiança) para outras que se aproximem mais de uma variável pivô. Nestas condições, é muito provável que o erro no nível de confiança de um intervalo baseado numa raiz "prepivoted" seja menor do que o erro no nível de confiança de um intervalo baseado na raiz inicial.

Atendendo aos diferentes pressupostos que presidiram à construção dos intervalos apresentados por Efron e à construção dos intervalos iterativos, torna-se difícil efectuar uma comparação teórica que leve a decidir por um dos intervalos, como sendo o melhor. Pode é adiantar-se que, para um exemplo

apresentado por Beran, o "Bootstrap" iterativo (com uma ou duas iterações) proporciona melhores intervalos de confiança do que os intervalos BC_a [veja-se Beran (1987) - pg. 463].

Dos cinco intervalos de confiança "Bootstrap" apresentados, é fácil de concluir que os baseados no método dos percentis e no método dos percentis corrigido do enviesamento são suplantados pelos intervalos BC_a. O problema reside na comparação entre os intervalos BC_a, os studentizados e os iterativos - é difícil dizer qual é o melhor, se é que existe, na verdade, um método melhor do que os outros.

O actual estado dos conhecimentos, nos domínios do "Bootstrap", de certa forma ainda embrionário, pelo menos nalguns aspectos, e a enorme diversidade de problemas concretos, levantam grandes dificuldades à dedução de leis gerais ou à construção de métodos que proporcionem melhores resultados em todas as situações. Os procedimentos a seguir e o método a utilizar dependem do caso concreto em estudo - a aplicação dos métodos "Bootstrap" é, e acredita-se que continuará a ser nos tempos mais próximos, uma aplicação mais casuísta (cada caso é um caso específico, adequando-se-lhe um determinado método) do que geral. Nesta linha de ideias, tem de se estar, forçosamente, de acordo com Dicicio e Romano, quando referem: « *In any given situation, the choice of bootstrap procedure depends on available theoretical results, computational considerations, the level of accuracy desired, simulation results and experience with similar problems. For example, both the BC_a and the percentile-t are second order correct; however, the BC_a requires knowledge of an analytical constant while the percentile-t requires a stable estimate of variance. Given the diversity of criteria in choosing a procedure, it is unlikely that a single procedure will emerge as a preferred method in all problems.* » [Dicicio e Romano (1988) - pg. 339].

4 - INTERVALOS DE CONFIANÇA "BOOTSTRAP" APLICADOS AO ÍNDICE DE GINI PARA OS RENDIMENTOS DOS PRODUTORES AGRÍCOLAS DOS AÇORES E DA MADEIRA

4.1 - O ÍNDICE DE GINI COMO UMA MEDIDA DE DESIGUALDADE

O estudo da desigualdade na distribuição do rendimento reveste-se de grande importância, quando se tem por objectivo analisar a vertente social de determinado sistema económico. Neste sentido, os economistas têm procurado (desde longa data) a construção de indicadores que permitam aferir o grau de desigualdade existente na distribuição dos rendimentos. De entre os indicadores propostos, merece especial destaque, pelos bons resultados obtidos e por ser, de longe, o mais utilizado em estudos sobre a desigualdade, o Índice de Gini.

O Índice de Gini pode ser utilizado para determinar o grau de concentração (logo, a desigualdade), não só dos rendimentos, como de qualquer variável, em relação à qual seja pertinente analisar a concentração. Como a aplicação tratada neste trabalho versa sobre os rendimentos dos produtores agrícolas, vai apresentar-se o Índice de Gini como uma medida estatística da desigualdade existente na distribuição dos rendimentos.

Considere-se, então, a seguinte variável aleatória,

X_i - rendimento do agregado (pode ser uma família ou um indivíduo) i.

Sabe-se que $X_i \sim F$, onde F representa a função de distribuição (desconhecida) da variável aleatória X_i. Repare-se que o rendimento de cada agregado, X_i, i = 1, 2, ..., obedece à mesma distribuição de probabilidade F.

O rendimento médio do agregado i é dado por,

$$\mu = E(X_i) = \int_0^{+\infty} x_i \; f(x_i) \; dx_i \, ,$$

onde, $f(x_i) \equiv \frac{dF(x_i)}{dx_i}$, é a função de densidade de probabilidade de X_i (está-se a supor, evidentemente, que X_i é uma variável aleatória contínua).

Defina-se,

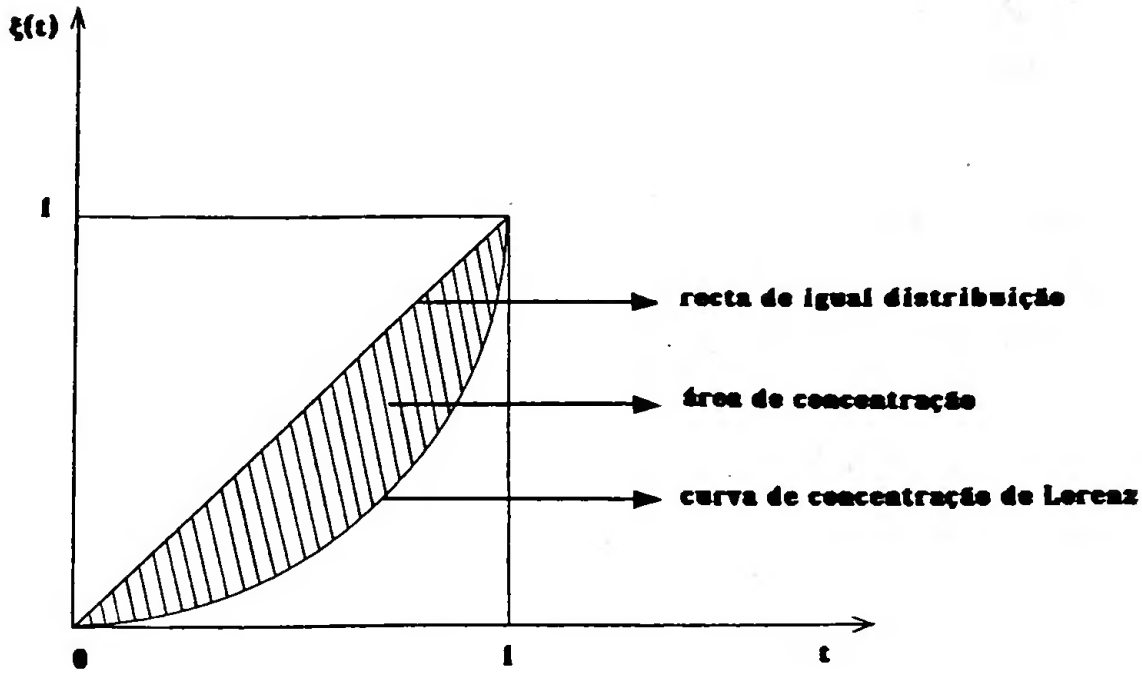
$$F^{-1}(t) = x_i : F(x_i) = t, \; 0 \leq t \leq 1.$$

Utilizando esta definição, pode construir-se,

$$\xi(t) = \frac{\int_0^{F^{-1}(t)} x_i \; f(x_i) \; dx_i}{\mu} = \frac{\int_0^{F^{-1}(t)} x_i \; f(x_i) \; dx_i}{\int_0^{+\infty} x_i \; f(x_i) \; dx_i} \, , \qquad (4.1)$$

função esta que é definida para $0 \leq t \leq 1$.

O gráfico da função $\xi(t)$, isto é, o gráfico de $\xi(t)$ em função de t, $0 \leq t \leq 1$, é a conhecida curva de concentração de Lorenz, cuja representação gráfica é dada por,



A área situada entre a curva de concentração de Lorenz e a recta (em rigor, segmento de recta) de igual distribuição dos rendimentos (sobre a qual a concentração é nula) é a chamada área de concentração. O Índice de Gini é dado pela razão entre a área de concentração e a área do triângulo onde esta se insere (repare-se que o Índice de Gini varia entre 0 - perfeita igualdade na distribuição dos rendimentos, isto é, concentração nula - e 1 - máxima concentração dos rendimentos). Como a área do triângulo é 1/2 (ver figura acima), o Índice de Gini acaba por ser o dobro da área de concentração.

Em termos da distribuição do universo, o Índice de Gini pode ser expresso como,

$$\theta \equiv \theta(F) \equiv G \equiv G(F) = \frac{E|X_i - X_j|}{2\,\mu} = \frac{\int_0^{+\infty}\int_0^{+\infty} |x_i - x_j|\, f(x_i)\, f(x_j)\, dx_i\, dx_j}{2\,\mu}. \qquad (4.2)$$

Tenha-se em atenção que as variáveis aleatórias X_i e X_j são independentes e idênticamente distribuídas, de acordo com a distribuição de probabilidade F.

Dada uma amostra casual $(X_1, X_2, ..., X_n)$, um estimador para $\theta \equiv G$ é,

$$\hat{\theta} \equiv \hat{\theta}_n \equiv \hat{\theta}(X_1, X_2, ..., X_n) \equiv \hat{G}(X_1, X_2, ..., X_n) = \frac{\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_i - X_j|}{2 \frac{1}{n} \sum_{i=1}^{n} X_i}, \qquad (4.3)$$

e a correspondente estimativa, tendo-se observado $(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$, é,

$$\hat{\theta} \equiv \hat{\theta}_n \equiv \hat{\theta}(x_1, x_2, ..., x_n) \equiv \hat{G}(x_1, x_2, ..., x_n) = \frac{\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2 \frac{1}{n} \sum_{i=1}^{n} x_i}. \qquad (4.4)$$

O Índice de Gini, que aqui foi sumariamente apresentado, é de difícil tratamento estatístico, na medida em que não se conhece a distribuição por amostragem da estatística $\hat{\theta}(X_1, X_2, ..., X_n) \equiv \hat{G}(X_1, X_2, ..., X_n)$, mesmo que se conheça a distribuição do universo. No caso presente, o problema é agravado, já que, nem tão-pouco se faz ideia sobre qual é a distribuição, F, do rendimento de cada agregado. Assim, encontramo-nos num domínio não paramétrico, o qual torna difícil as comparações estatísticas entre o grau de concentração de universos diferentes. Está-se, pois, em presença de um problema de inferência em que o "Bootstrap" parece ter "uma palavra a dizer", permitindo ultrapassar as restrições colocadas pela escassa informação existente (apenas as concretas amostras observadas).

4.2 - A CONSTRUÇÃO DAS AMOSTRAS DOS RENDIMENTOS DOS PRODUTORES AGRÍCOLAS DOS AÇORES E DA MADEIRA

Os dados analisados no presente estudo dizem respeito aos rendimentos dos produtores agrícolas dos Açores e da Madeira, tendo sido recolhidos no âmbito do Inquérito às Receitas e Despesas Familiares 1980/81 (IRDF), levado a cabo pelo Instituto Nacional de Estatística (INE).

O IRDF teve a sua fase de inquirição entre Março de 1980 e Fevereiro de 1981, abrangendo geograficamente o Continente, a Região Autónoma da Madeira e a Região Autónoma dos Açores, embora nesta última se circunscrevesse apenas à ilha de São Miguel, devido às ocorrências sísmicas de 1 de Janeiro de 1980.

A unidade estatística central do IRDF é o "Agregado Doméstico Privado", tendo-se recolhido variados dados sócio-económicos, de entre os quais as receitas dos indivíduos representantes do "Agregado Doméstico Privado". São precisamente estas receitas, consideradas para a classe sócio-económica dos produtores agrícolas e para as regiões autónomas dos Açores e da Madeira, que constituem as duas amostras concretas (a dos Açores com 143 valores de receita e a da Madeira com 163) que vão permitir a construção dos intervalos de confiança "Bootstrap" para os Índices de Gini aplicados aos rendimentos dos produtores agrícolas dos Açores e da Madeira.

Os valores da receita recolhidos dizem respeito às receitas auferidas pelos respectivos titulares nos doze meses anteriores ao mês da entrevista e englobam:

- As receitas monetárias ordinárias, como sejam, as receitas provenientes do trabalho por conta de outrém e por conta própria, os rendimentos das empresas em nome individual incluídas nas Famílias, os levantamentos ou rendimentos de empresas quase--sociedades, os rendimentos monetários de membros activos de cooperativas de produção, os rendimentos de propriedade e as transferências regulares ou periódicas.

- As receitas monetárias extraordinárias, como é o caso das transferências não periódicas.

- As receitas em espécie ou natureza, como sejam, o autoconsumo, o autoabastecimento, a autolocação, as remunerações em natureza provenientes do trabalho por conta de outrém, os bens e serviços fornecidos gratuitamente ou a preços reduzidos e os rendimentos em natureza de membros activos das cooperativas de produção.

Finalmente, refira-se que a selecção das unidades de alojamento inquiridas foi realizada segundo o método de amostragem probabilística e multietápica.

4.3 - ALGUMAS CONSIDERAÇÕES SOBRE A APLICAÇÃO DOS INTERVALOS DE CONFIANÇA "BOOTSTRAP" ÀS CONCRETAS AMOSTRAS OBSERVADAS

Os intervalos de confiança "Bootstrap", deduzidos no ponto 3 do presente trabalho, vão ser aplicados às concretas amostras de rendimentos dos produtores agrícolas dos Açores e da Madeira, por forma a construir aproximações aos exactos intervalos de confiança a $(1-2\alpha)100\%$ para os Índices de Gini referentes aos citados rendimentos.

Consideraram-se dois níveis de confiança: 90% (fazendo $\alpha = 0.05$) e 95% (fazendo $\alpha = 0.025$).

Atendendo a que foram deduzidos, teoricamente, cinco intervalos de confiança "Bootstrap", vão construir-se vinte intervalos de confiança: dez para o Índice de Gini referente aos rendimentos dos produtores agrícolas dos Açores (os cinco deduzidos teoricamente a 90% e a 95%); dez para o Índice de Gini referente aos rendimentos dos produtores agrícolas da Madeira (os cinco deduzidos teoricamente a 90% e a 95%).

Os intervalos de confiança vão ser construídos com base nas fórmulas explanadas no ponto 3 deste trabalho, isto é:

- Fórmulas (3.13), (3.14) e (3.15), para o método dos percentis (designado abreviadamente por PER).

- Fórmulas (3.61), (3.62) e (3.63), para o método dos percentis corrigido do enviesamento (designado abreviadamente por BC).

- Fórmulas (3.90), (3.91) e (3.92), para o método dos percentis corrigido do enviesamento e da aceleração da variância (designado abreviadamente por BC_a).

- Fórmulas (3.106), (3.107) e (3.108), para o "Bootstrap - t" (designado abreviadamente por Bt).

- Fórmulas (3.114), (3.123) e (3.124), para o "Bootstrap" iterativo (designado abreviadamente por IT). No caso do "Bootstrap" iterativo, apenas se realiza uma iteração (intervalos $B_{n,1}$). A raiz escolhida para o intervalo de confiança é,

$$R \equiv R_n(\theta) \equiv R_n(X_1, X_2, ..., X_n ; \theta) = \frac{\hat{\theta}_n - \theta}{\hat{\sigma}_{\hat{\theta}}} ,$$

onde $\hat{\theta}_n$ é dado por (4.3) e θ por (4.2), sendo $\hat{\sigma}_{\hat{\theta}}$ um estimador consistente do desvio padrão de $\hat{\theta}_n$ (note-se que esta raiz não é mais do que a variável aleatória T utilizada no "Bootstrap - t"). Os valores "Bootstrap" desta raiz (utilizados na estimação de $\hat{J}_n$) são dados por,

$$R_n(\underline{X}_n^{*b}, \hat{\theta}_n) = \frac{\hat{\theta}_n^{*b} - \hat{\theta}_n}{\hat{\sigma}_{BOOT}} , \; b = 1, 2, ..., B_1,$$

onde $\hat{\theta}_n^{*b} \equiv \hat{\theta}(\underline{X}_n^{*b})$, $b = 1, 2, ..., B_1$, $\hat{\theta}_n$ é dado por (4.4) e $\hat{\sigma}_{BOOT}$ é dado por (2.9) e (2.8), fazendo $B = B_1$ nestas fórmulas.

Os valores "Bootstrap" de "segunda ordem" da raiz escolhida (utilizados na estimação de $\hat{J}_{n,1}$) são dados por,

$$R_n(\underline{X}^{**t}_{n,b}, \hat{\theta}^{*b}_n) = \frac{\hat{\theta}^{**t}_{n,b} - \hat{\theta}^{*b}_n}{\hat{\sigma}^{**}_b} \quad , b = 1, 2, ..., B_1, t = 1, 2, ..., B_2,$$

onde $\hat{\theta}^{**t}_{n,b} \equiv \hat{\theta}(\underline{X}^{**t}_{n,b})$, $b = 1, 2, ..., B_1$, $t = 1, 2, ..., B_2$, e

$$\hat{\sigma}^{**}_b = \sqrt{[1/(B_2-1)] \sum_{t=1}^{B_2} (\hat{\theta}^{**t}_{n,b} - \hat{\mu}^{**}_b)^2} \quad , \quad \hat{\mu}^{**}_b = \sum_{t=1}^{B_2} (1/B_2)\ \hat{\theta}^{**t}_{n,b} \quad .$$

Quanto ao número de réplicas utilizado para construir os intervalos de confiança, optou-se por efectuar 1000 réplicas, como é aconselhado por alguns autores [ver nota de pé-de-página nº 9]. No "Bootstrap" iterativo fez-se $B_1 = B_2 = 1000$, ou seja, o número total de réplicas "Bootstrap" é de $B_1 + B_1\ B_2 = 1001000$, sendo 1000 de "primeira ordem" e 1000000 de "segunda ordem".

O programa informático que implementou a construção dos intervalos de confiança "Bootstrap" (do qual se junta uma listagem em anexo [ver Anexo 2]) foi construído em linguagem Pascal e correu num computador MICROVAX 3600.

4.4 - ANÁLISE DOS RESULTADOS EMPÍRICOS OBTIDOS

Os valores do Índice de Gini para as particulares amostras dos rendimentos dos produtores agrícolas dos Açores e da Madeira, que serviram de base a este estudo empírico, são muito semelhantes, apresentando a amostra da Madeira uma concentração dos rendimentos ligeiramente mais elevada do que a amostra dos Açores:

QUADRO 1

	ÍNDICE DE GINI - $\hat{\theta}(X_1, X_2, ..., X_n)$
Açores	0.400141
Madeira	0.409430

Para a construção dos intervalos BC e BC_a , torna-se necessário calcular as constantes z_0 e "a".

A constante de enviesamento, z_0 , é determinada (aproximadamente) através da expressão (3.53), a qual envolve a estimativa da função de distribuição empírica "Bootstrap" de $\hat{\theta}(X_1, X_2, ..., X_n)$, isto é, $\hat{G}(t)$. Por sua vez, esta última é dada por (3.12), encontrando-se o seu gráfico nas Figuras 1 (para os Açores) e 2 (para a Madeira) [ver Anexo 1]. Cabe aqui realçar a grande regularidade das duas funções de distribuição empíricas (obtidas com 1000 réplicas "Bootstrap" de $\hat{\theta}(X_1, X_2, ..., X_n)$), encontrando-se (graficamente) grandes

semelhanças com a função de distribuição da normal. Aliás, os valores do coeficiente de assimetria e do kurtosis das distribuições empíricas "Bootstrap" de $\hat{\theta}(X_1, X_2, ..., X_n)$, mostram-se muito semelhantes aos valores conhecidos da distribuição normal (que são zero, quer para o coeficiente de assimetria, quer para o kurtosis):

QUADRO 2

	ESTIMATIVA DO COEFICENTE DE ASSIMETRIA DE $\hat{\theta}(X_1, X_2, ..., X_n)$	ESTIMATIVA DO KURTOSIS DE $\hat{\theta}(X_1, X_2, ..., X_n)$
Açores	-0.052015	-0.119332
Madeira	-0.096813	-0.144416

O cálculo da constante z_0 acabou por dar os seguintes valores:

QUADRO 3

	z_0
Açores	0.095396
Madeira	0.123135

Observa-se um enviesamento maior para o caso da Madeira, embora a diferença não seja muito acentuada em relação ao caso dos Açores.

Quanto à constante de acelaração "a", esta é determinada (aproximadamente) através da expressão (3.96). Esta última, envolve a

determinação da função de influência empírica, dada pelas condições (3.97), (3.98) e (3.99), cujo gráfico se encontra nas figuras 3 (para os Açores) e 4 (para a Madeira) [ver Anexo 1]. Os gráficos das funções de influência empíricas mostram a grande contribuição dos rendimentos muito baixos e muito elevados para a concentração dos rendimentos, sendo de realçar o maior peso dos rendimentos muito elevados para a concentração atrás referida.

Os valores obtidos para a constante "a" foram os seguintes:

QUADRO 4

	a
Açores	0.064259
Madeira	0.040235

Os concretos intervalos de confiança "Bootstrap" para os rendimentos dos produtores agrícolas dos Açores e da Madeira encontram-se nos Quadros 5 e 6:

QUADRO 5

	INTERVALOS DE CONFIANÇA "BOOTSTRAP" - AÇORES			
	90%		95%	
	LIM. INFERIOR	LIM. SUPERIOR	LIM. INFERIOR	LIM. SUPERIOR
PER	0.3465	0.4456	0.3377	0.4534
BC	0.3523	0.4498	0.3417	0.4609
BC_a	0.3573	0.4564	0.3481	0.4753
Bt	0.3546	0.4537	0.3468	0.4624
IT	0.3546	0.4564	0.3483	0.4624

QUADRO 6

	INTERVALOS DE CONFIANÇA "BOOTSTRAP" - MADEIRA			
	90%		95%	
	LIM. INFERIOR	LIM. SUPERIOR	LIM. INFERIOR	LIM. SUPERIOR
PER	0.3629	0.4445	0.3569	0.4533
BC	0.3694	0.4507	0.3619	0.4609
BC_a	0.3729	0.4559	0.3636	0.4631
Bt	0.3743	0.4558	0.3564	0.4619
IT	0.3770	0.4567	0.3667	0.4635

Da análise dos quadros atrás apresentados, conclui-se que os intervalos de confiança construídos segundo os cinco métodos "Bootstrap" apresentados

acabam por não ser muito distintos, o que poderá ser explicado pelo baixo valor das constantes de enviesamento e de aceleração da variância.

Como era de esperar, os intervalos de confiança não são simétricos em torno de $\hat{\theta}(X_1, X_2, ..., X_n)$, podendo observar-se melhor esta assimetria através da seguinte relação,

$$D/E = \frac{\text{Extremo superior do intervalo de confiança} - \hat{\theta}(x_1, x_2, ..., x_n)}{\hat{\theta}(x_1, x_2, ..., x_n) - \text{Extremo inferior do intervalo de confiança}},$$

cujos valores se encontram nos Quadros 7 e 8:

QUADRO 7

	RELAÇÃO D/E PARA OS INTERVALOS DE CONFIANÇA A 90%				
	PER	BC	BC_a	Bt	IT
Açores	0.85	1.04	1.31	1.18	1.24
Madeira	0.75	1.03	1.27	1.32	1.46

QUADRO 8

	RELAÇÃO D/E PARA OS INTERVALOS DE CONFIANÇA A 95%				
	PER	BC	BC_a	Bt	IT
Açores	0.85	1.04	1.44	1.17	1.20
Madeira	0.84	1.08	1.17	1.19	1.27

Nos métodos BC, Bt e IT, os intervalos de confiança são mais assimétricos no caso da Madeira do que no caso dos Açores (exceptua-se o caso dos intervalos BC a 90%, em que a assimetria é maior nos Açores), o que se deve ao maior valor da constante de enviesamento na primeira região. No entanto, nos intervalos BC_a, a assimetria é significativamente superior no caso dos Açores, o que se deve ao maior valor da constante de aceleração da variância para esta região - parece lícito tirar a conclusão de que a maior aceleração da variância para os Açores mais do que compensa o maior enviesamento para a Madeira, acabando por tornar os intervalos BC_a mais assimétricos no caso dos Açores.

Nos Quadros 9 e 10, podem ver-se as amplitudes dos intervalos de confiança construídos:

QUADRO 9

	AMPLITUDES DOS INTERVALOS DE CONFIANÇA A 90%				
	PER	BC	BC_a	Bt	IT
Açores	0.0991	0.0975	0.0991	0.0991	0.1018
Madeira	0.0816	0.0813	0.0830	0.0815	0.0797

QUADRO 10

	AMPLITUDES DOS INTERVALOS DE CONFIANÇA A 95%				
	PER	BC	BC_a	Bt	IT
Açores	0.1157	0.1192	0.1272	0.1156	0.1141
Madeira	0.0964	0.0990	0.0995	0.0965	0.0968

As amplitudes dos diferentes intervalos de confiança acabam por ser semelhantes, sendo, no entanto, de destacar o método iterativo como o que dá menores amplitudes em dois dos quatro casos: intervalos de confiança a 90% para a Região da Madeira e intervalos de confiança a 95% para a Região dos Açores. Além disso, também no caso dos intervalos de confiança a 95% para a Região da Madeira, o método iterativo gera um intervalo de confiança com amplitude perto da amplitude mínima verificada.

A construção destes intervalos de confiança "Bootstrap" permite-nos identificar um leque de valores, entre os quais se encontra o "verdadeiro" Índice de Gini, o que abre outras perspectivas e hipóteses de comparação quanto à concentração dos rendimentos dos produtores agrícolas nas regiões dos Açores e da Madeira. Efectivamente, dizer que a concentração atrás referida é maior na Madeira do que nos Açores, tendo por base apenas os valores amostrais do Índice de Gini (40.9%, para a Madeira e 40.0%, para os Açores), equivale a fazer um raciocínio muito superficial e sem grandes fundamentos. Os valores da amostra nem sempre são um "espelho" dos valores do universo, daí que se torne necessário trabalhar a amostra para, a partir dos valores que ela proporciona, inferir sobre o universo em causa. Esta metodologia básica e fundamental da estatística torna-se difícil de aplicar em situações de escassa informação, nomeadamente, em domínios não paramétricos, problema que é ultrapassado, como já se viu, pelo "Bootstrap". Na presente aplicação empírica, a construção dos intervalos de confiança "Bootstrap" para os rendimentos dos produtores agrícolas dos Açores e da Madeira, permite ver como os intervalos de confiança para as duas regiões têm um significativo conjunto de valores em comum, pelo que, em termos do "verdadeiro" Índice de Gini, a maior concentração dos rendimentos dos produtores agrícolas tanto poderá ser na Madeira como nos Açores e com

probabilidades consideráveis em ambos os casos, embora sempre mais favoráveis ao caso da Madeira.

5 - CONCLUSÕES

O "Bootstrap" veio abrir novos horizontes à inferência estatística em domínios não paramétricos. No campo da construção de intervalos de confiança, a contribuição do "Bootstrap" é muito importante, tendo-se deduzido vários intervalos de confiança, a partir de diferentes hipóteses e/ou perspectivas iniciais.

Problemas não paramétricos, até há pouco sem solução, ou com solução dada por aproximações assintóticas algo forçadas e sem grande justificação teórica, passam a ser resolvidos pelo "Bootstrap", com alguma simplicidade e elegância.

A grande lacuna do "Bootstrap" advém do seu aparecimento ainda algo recente (pode considerar-se que o "Bootstrap" nasceu em 1979, com Efron), o que justifica a inexistência de um corpo teórico unificado e coerente. Assim, está em aberto a investigação, nomeadamente, no que diz respeito à teoria assintótica do "Bootstrap", permitindo a dedução de propriedades e a comparação estatística de diferentes procedimentos "Bootstrap", ou no que concerne a estudos de sensibilidade face a diferentes amostras concretas observadas (um dos problemas, por vezes esquecido na aplicação do "Bootstrap", é o de se poder estar a fazer as reamostragens a partir de uma amostra "pouco representativa" do universo que a originou).

Neste trabalho, as lacunas teóricas atrás referidas sentiram-se nas dificuldades em comparar os cinco intervalos de confiança "Bootstrap"

apresentados, em especial, os intervalos BC_a , Bt e IT. Como se disse, no ponto 3.6, a aplicação do "Bootstrap" à construção de intervalos de confiança não paramétricos reveste-se de um carácter mais casuísta do que geral, estando-se, ainda, algo distante de um método unificado que proporcione a melhor solução em todos os problemas concretos.

Não obstante estas lacunas (que a constante e insistente investigação promete extinguir num futuro próximo) é de realçar a valia do "Bootstrap", permitindo apresentar um leque de intervalos de confiança, em áreas onde ainda tal não fora possível, como é o caso do Índice de Gini. A aplicação realizada neste trabalho foi mais além do simples valor amostral do Índice de Gini (até aqui, a única informação apreendida pelos economistas), construindo-se alguns intervalos de confiança "Bootstrap" para os "verdadeiros" Índices de Gini dos rendimentos dos produtores agrícolas dos Açores e da Madeira, o que possibilita uma mais correcta comparação da concentração dos rendimentos nas regiões autónomas atrás referidas (que se verificou estar situada dentro de limites quase iguais para as duas regiões).

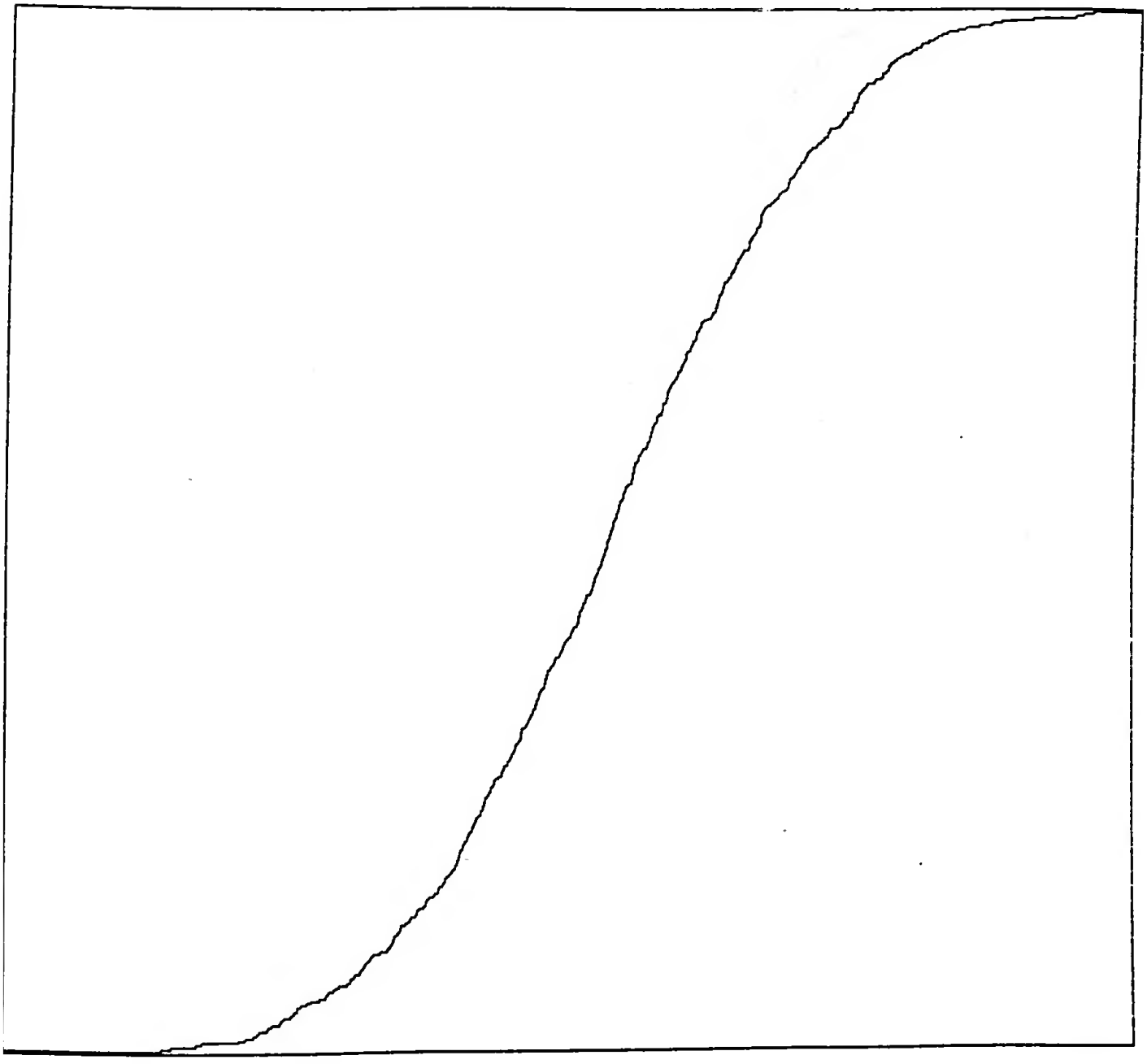
Em suma, o "Bootstrap" é uma teoria que dá capacidade de resposta para muitos problemas não paramétricos, em que a escassez de informação é acentuada e se circunscreve, praticamente, à concreta amostra observada, não devendo, no entanto, ser encarado como uma solução definitiva, mas, pelo contrário, como um promissor caminho que há pouco começou a ser percorrido e sobre o qual ainda se terá de trilhar muito até alcançar respostas mais globais e unificadoras.

6 - ANEXOS

ANEXO 1 - GRÁFICOS DAS FUNÇÕES DE DISTRIBUIÇÃO EMPÍRICAS "BOOTSTRAP" E DAS FUNÇÕES DE INFLUÊNCIA EMPÍRICAS DO ESTIMADOR DOS ÍNDICES DE GINI PARA OS RENDIMENTOS DOS PRODUTORES AGRÍCOLAS DOS AÇORES E DA MADEIRA E DAS FUNÇÕES DE DISTRIBUIÇÃO EMPÍRICAS "BOOTSTRAP" DA VARIÁVEL ALEATÓRIA UTILIZADA COMO RAIZ ORIGINAL NO "BOOTSTRAP" ITERATIVO

FIGURA 1

FUNÇÃO DE DISTRIBUIÇÃO EMPÍRICA "BOOTSTRAP" DO ESTIMADOR DO ÍNDICE DE GINI PARA OS RENDIMENTOS DOS PRODUTORES AGRÍCOLAS DOS AÇORES

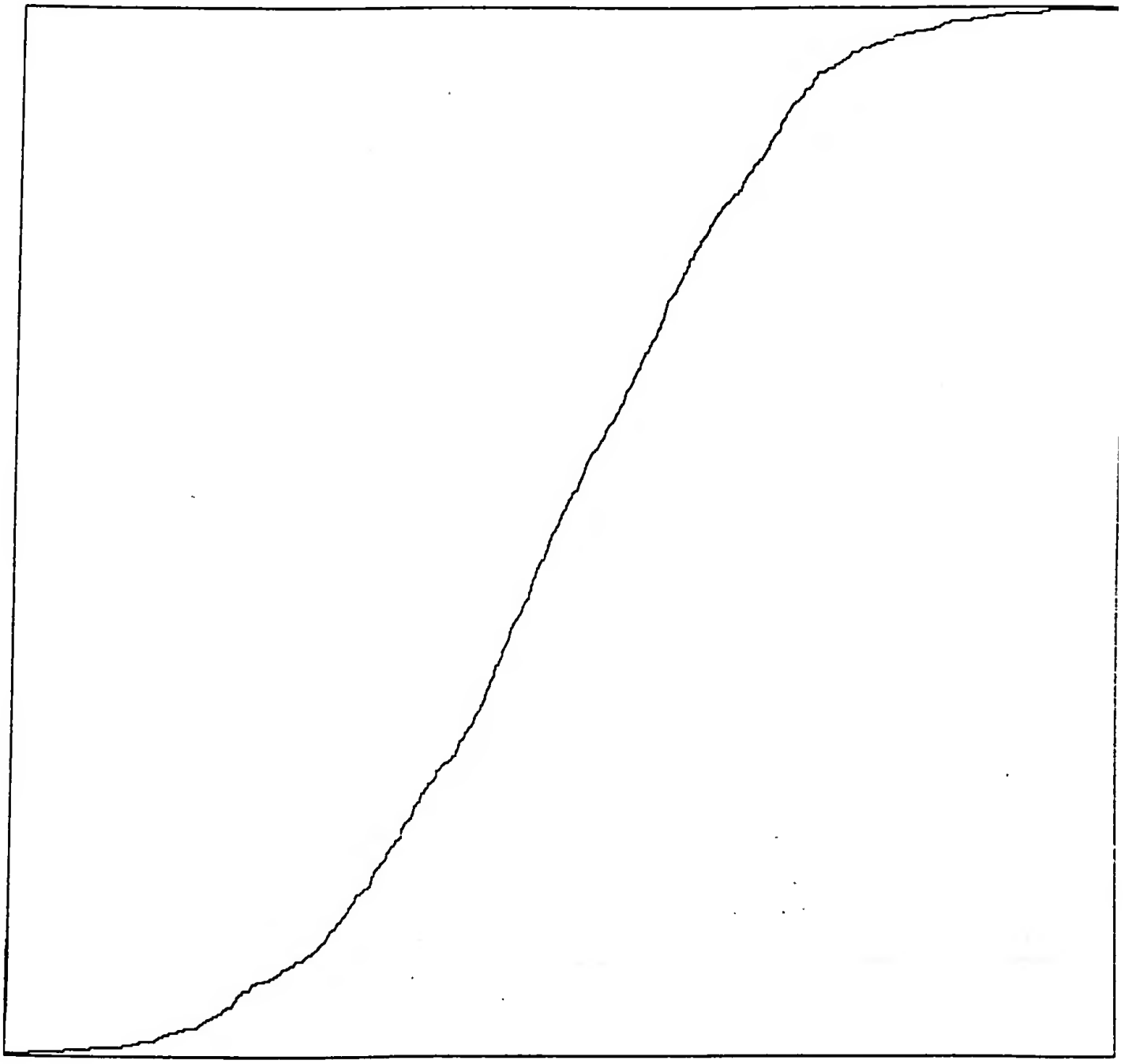


	min	max
X	2.88355056E-01	4.86789558E-01
Y	0.00000000E+00	1.00000000E+00

Razão entre escalas (Y/X) 5.04445560E+00

FIGURA 2

FUNÇÃO DE DISTRIBUIÇÃO EMPÍRICA "BOOTSTRAP" DO ESTIMADOR DO ÍNDICE DE GINI PARA OS RENDIMENTOS DOS PRODUTORES AGRÍCOLAS DA MADEIRA

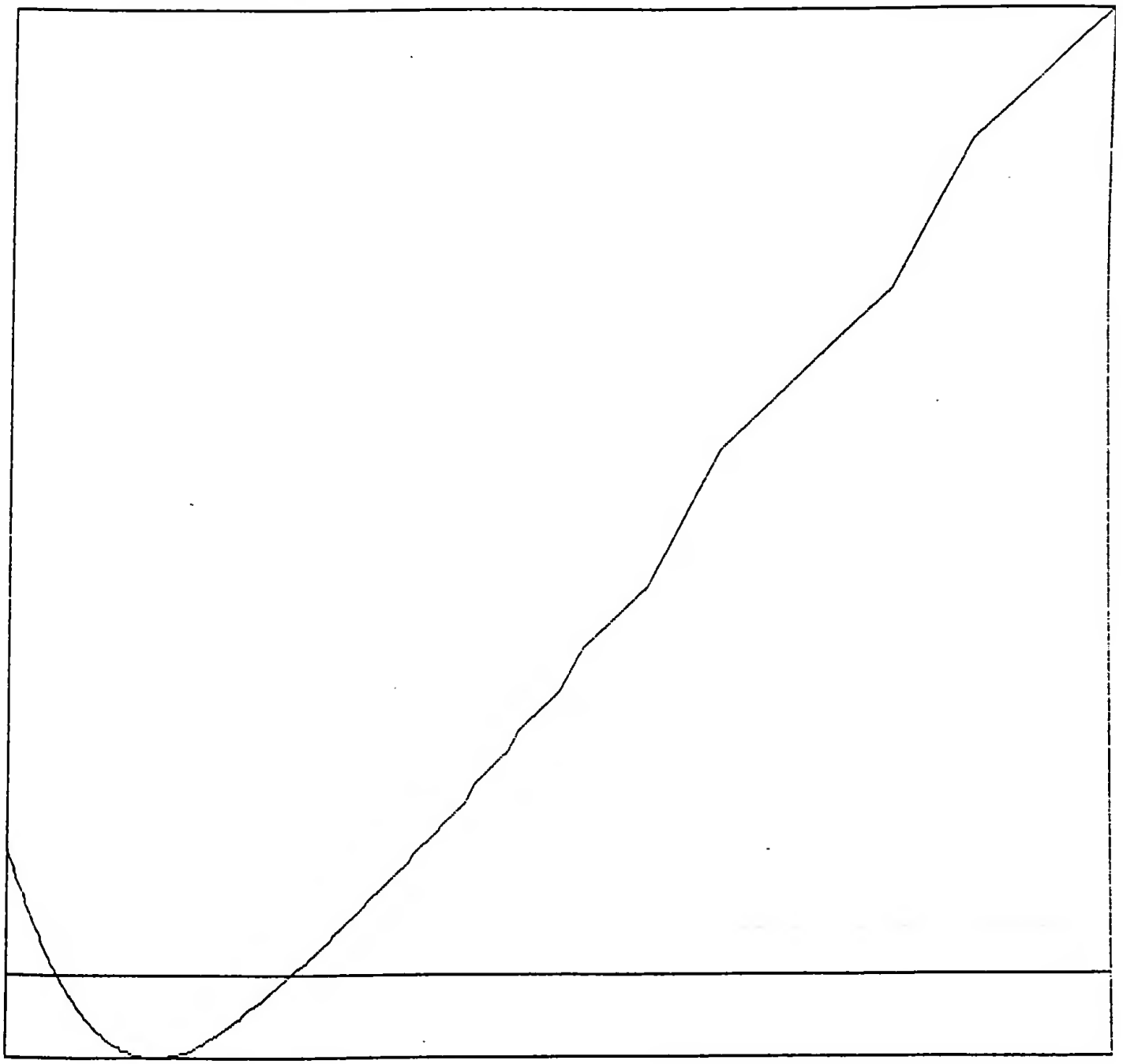


	min	max
X	3.31496286E-01	4.82668203E-01
Y	0.00000000E+00	1.00000000E+00

Razão entre escalas (Y/X) 6.62156081E+00

FIGURA 3

FUNÇÃO DE INFLUÊNCIA EMPÍRICA DO ESTIMADOR DO ÍNDICE DE GINI PARA OS RENDIMENTOS DOS PRODUTORES AGRÍCOLAS DOS AÇORES

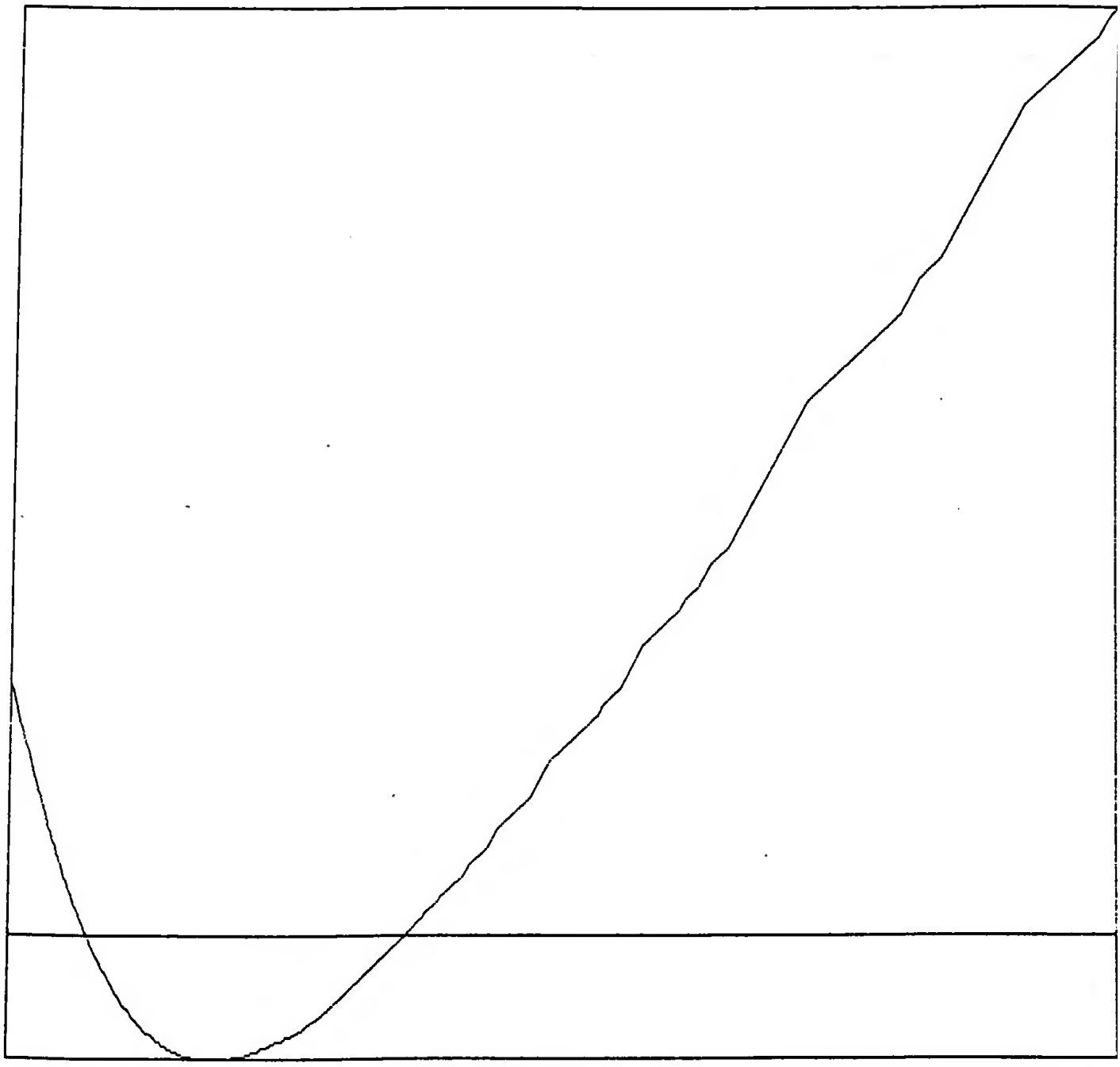


	min	max
X	3.23742600E+04	1.36706480E+06
Y	-2.42466624E-01	2.90491735E+00

Razão entre escalas (Y/X) 2.36048169E-06

FIGURA 4

FUNÇÃO DE INFLUÊNCIA EMPÍRICA DO ESTIMADOR DO ÍNDICE DE GINI PARA OS RENDIMENTOS DOS PRODUTORES AGRÍCOLAS DA MADEIRA

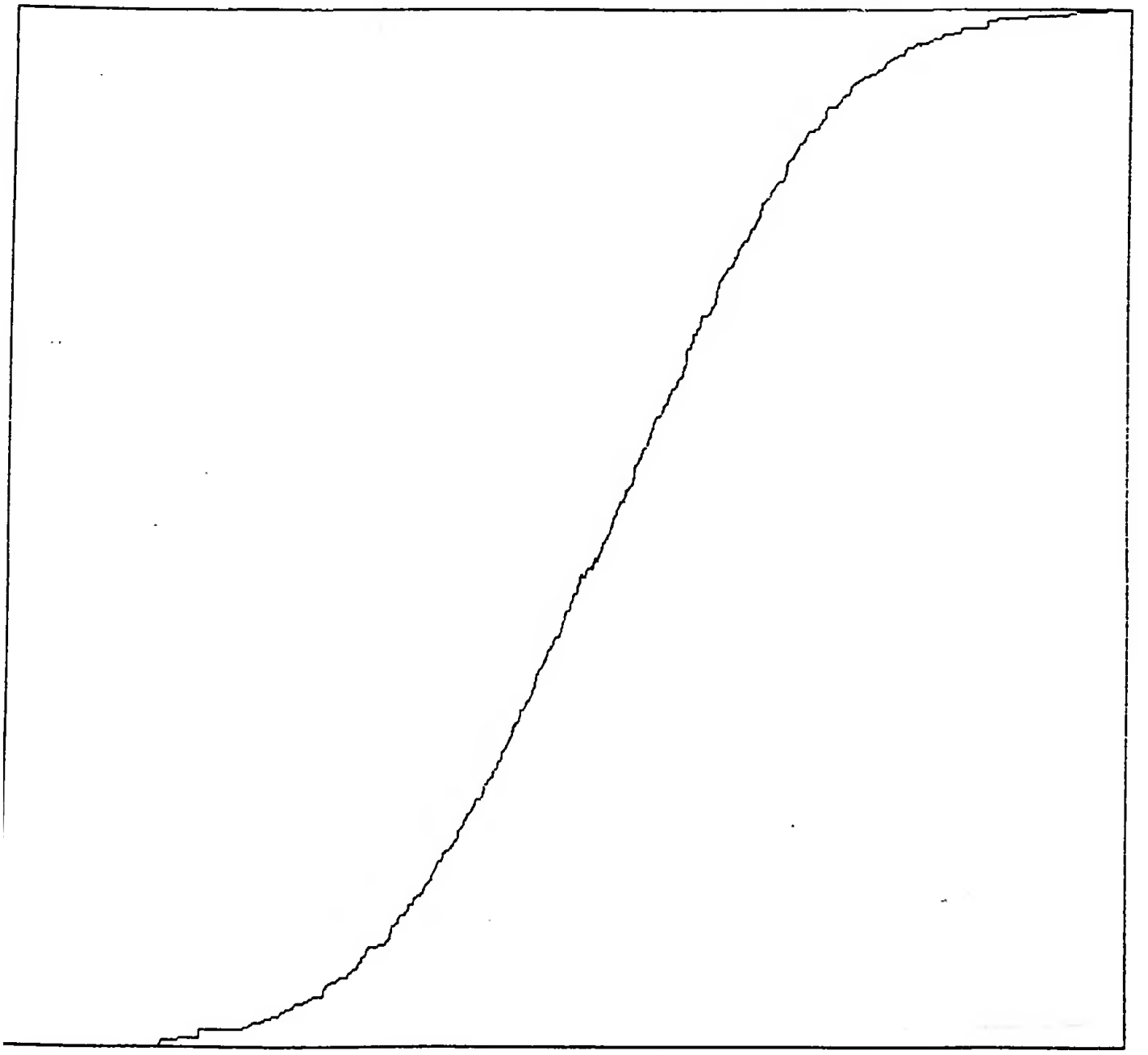


	min	max
X	1.17930000E+04	8.62151000E+05
Y	-2.35421160E-01	1.82334297E+00

Razão entre escalas (Y/X) 2.42346237E-06

FIGURA 5

FUNÇÃO DE DISTRIBUIÇÃO EMPÍRICA "BOOTSTRAP" DA VARIÁVEL ALEATÓRIA UTILIZADA COMO RAIZ ORIGINAL NO "BOOTSTRAP" ITERATIVO - RENDIMENTOS DOS PRODUTORES AGRÍCOLAS DOS AÇORES

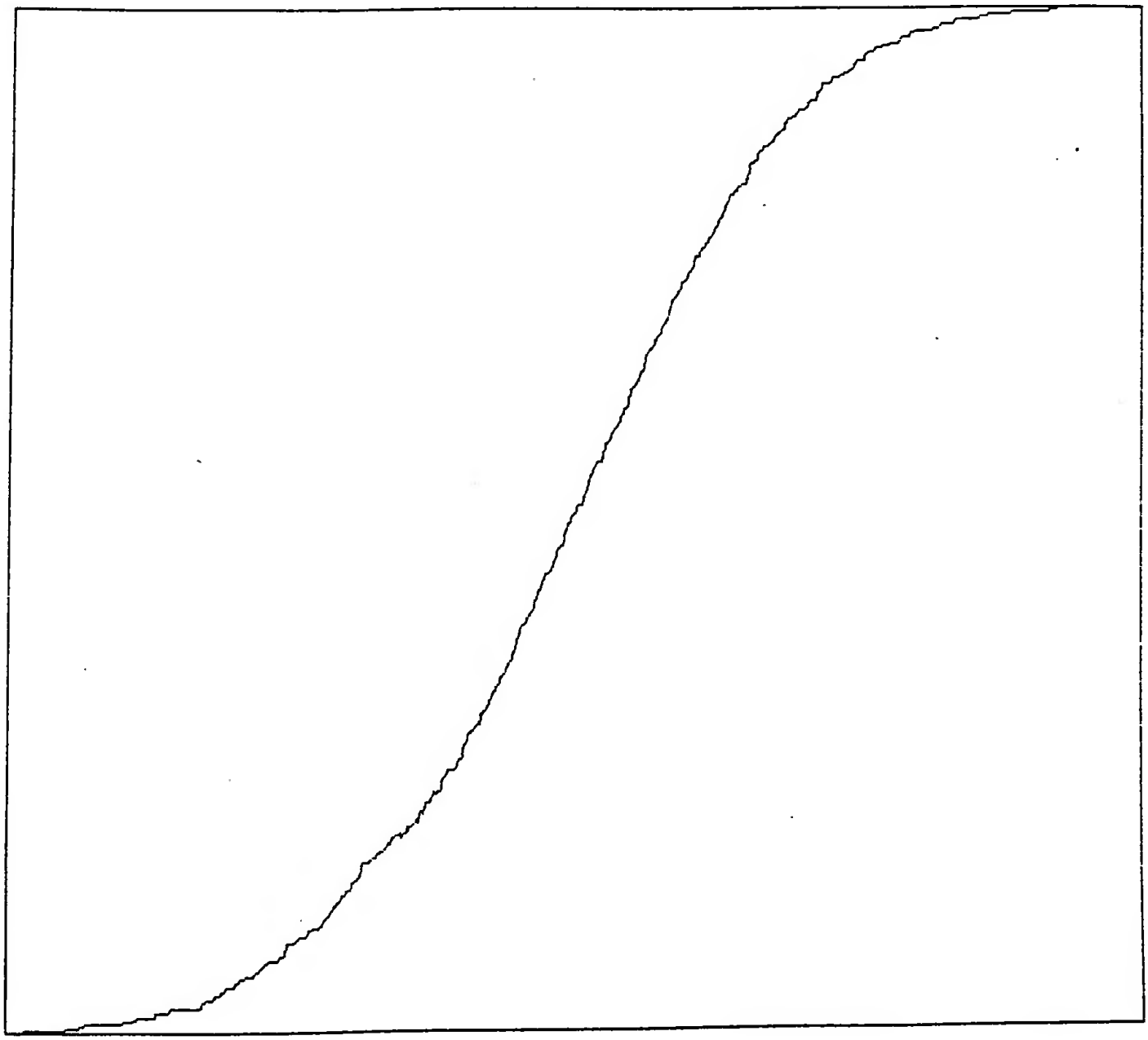


	min	max
(	-3.71579365E+00	2.88020049E+00
'	0.00000000E+00	1.00000000E+00

azão entre escalas (Y/X) 1.51757872E-01

FIGURA 6

FUNÇÃO DE DISTRIBUIÇÃO EMPÍRICA "BOOTSTRAP" DA VARIÁVEL ALEATÓRIA UTILIZADA COMO RAIZ ORIGINAL NO "BOOTSTRAP" ITERATIVO - RENDIMENTOS DOS PRODUTORES AGRÍCOLAS DA MADEIRA



	min	max
X	-3.13566975E+00	2.94672824E+00
Y	1.00000005E-03	1.00000000E+00

Razão entre escalas (Y/X) 1.64407696E-01

ANEXO 2 - LISTAGEM DO PROGRAMA ELABORADO EM LINGUAGEM PASCAL PARA A CONSTRUÇÃO DOS INTERVALOS DE CONFIANÇA "BOOTSTRAP"

Nota: o programa cuja listagem se segue foi construído pelo Dr. João Manuel Andrade e Silva, tendo por base uma versão original (versão 1.0), também de sua autoria, utilizada no estudo empírico realizado em Murteira (1988b), a qual foi publicada nesse mesmo trabalho.

```
Program BOOTSTRAP (input,output);
    {
        Este programa destina-se a efectuar o estudo "bootstrap" de uma
      estatística extraída de uma amostra de dimensão M de uma população
      com distribuição desconhecida.
        Para tal, tirar-se-ão N amostras com reposição da amostra atrás
      referida.
        A constante M_N_MAX destina-se a definir a dimensão máxima, que
      da amostra inicial, quer do número de amostras que se vão extrair
      no estudo "bootstrap", isto é, M_N_MAX constitui um limite superio
      para max(M,N).
        Para correr o programa torna-se necessário definir um ficheiro
      contendo os valores da amostra inicial. Tal ficheiro deve conter
      um valor por linha e não deve conter "linhas em branco" nem linhas
      de comentários. O valor M é definido pela contagem das linhas do
      ficheiro.
        A nível de input, também se vai inquirir de forma interactiva o
      número de amostras a tirar com reposição, a partir da amostra
      inicial (N) bem como o valor gerador dos números pseudo aleatórios
        O output é dirigido para 7 ficheiros
            RES.DAT        -- contém os resultados mais importantes.
                              Formatado a 80 colunas para ser impresso.

            GRAFICO1.DAT   -- coordenadas dos pontos (Xi,Ui) por ordem
                              crescente dos Xi, por forma a correr o
                              programa GRAFICO.

            GRAFICO2.DAT   -- coordenadas dos pontos de salto da
                              distribuição amostral de TETA por ordem
                              crescente de TETA, por forma a correr o
                              programa GRAFICO.

            HISTO.DAT      -- histograma dos valores de TETA calculado
                              desprezando a observação i (M valores).
                              O histograma é construído com classes de
                              amplitude 0.02

            VEC_TETA.DAT   -- valores de TETA nas condições anteriores
                              ordenados por ordem crescente.

            VECTORES.DAT   -- coordenadas dos pontos de salto da distri-
                              buição base do "bootstrap" iterativo.

            ALEATORIO.DAT-- contém informações sobre os números
                              aleatórios gerados pelo programa

      versão 1.1 ----   Janeiro 1990 ---- João Manuel Andrade e Silva

}

CONST
    M_N_MAX  =  1000;

TYPE
    VECTOR1 = array [1..M_N_MAX] of double;
    VECTOR2 = array [1..M_N_MAX] of integer;
    MATRIZ  = array [1..M_N_MAX,1..M_N_MAX] of double;
    NOME_TE = packed array [1..18] of char;

VAR
    OBS,AMOSTRA,TETA_    : vector1;
    CONTADOR,CONT        : vector2;
    MAT_TETA             : matriz;
    NOME_TETA            : nome_te;
    N1,N2,M,I            : integer;
```

```
    SAI                     : boolean;
    SEED                    : unsigned;
    SAIDA                   : text;
    TETA,A,T1,T2            : double;

{############## BLOCO DAS FUNÇÕES E PROCEDIMENTOS EXTERNOS ########}

Function CALC_TETA (X:vector1; DIMENSAO:integer):double;extern;
        {Calcula a estatística que vai ser objecto do estudo}


Procedure NOME (var NOME_TETA:nome_te);extern;
        {Define o nome da estatística em estudo}


{############## BLOCO DAS FUNÇÕES E PROCEDIMENTOS GERAIS ########}


Procedure SWAP(var A,B:double);
        {Troca os valores de A e B}
    var
        C:double;
    Begin
    C:=A; A:=B; B:=C;
    End; {procedure swap}


Procedure ORDENA (var X:VECTOR1; DIMENSAO:integer);
        {Devolve o vector X com os seus DIMENSAO primeiros
         elementos ordenados por ordem crescente.
         Utiliza o procedimento Swap.
         Ter em atenção que o procedimento é pouco eficiente}
    VAR
        I,J : integer;
    Begin
    for I:=1 to DIMENSAO-1 do
        for J:=I+1 to DIMENSAO do
            if (X[i]>X[j]) then swap(X[i],X[j]);
    End; {Procedure ORDENA}


Procedure ORDENA1 (var X:matriz; L,DIMENSAO:integer);
        {Devolve a linha L da matriz X com os seus DIMENSAO primeiros
         elementos ordenados por ordem crescente.
         Utiliza o procedimento Swap.
         Ter em atenção que o procedimento é pouco eficiente}
    VAR
        I,J : integer;
    Begin
    for I:=1 to DIMENSAO-1 do
        for J:=I+1 to DIMENSAO do
            if (X[l,i]>X[l,j]) then swap(X[l,i],X[l,j]);
    End; {Procedure ORDENA}



Function NORMAL (X:double):double;
        {Calcula a probabilidade de valores < ou = a X}
    var
        A,B,C,D : double; N : integer;
    begin
    if (abs(X)>4.5)
            then
```

```
        if (X>0) then NORMAL:=1 else NORMAL:=0
            else
        if (abs(X))>3.7) then
            begin
            A := 1/sqrt(8*arctan(1));
            B := exp(-0.5*X*X)*A/abs(X);
            C := 1-1/(X*X)+3/(X**4)-15/(X**6)+105/(X**8);
            if (X>0) then NORMAL:=1-B*C else NORMAL:=B*C;
            end
                        else
            begin
            A := 1/sqrt(8*arctan(1));
            B := 0.5 + A*X - A*X*X*X/6.0;
            C := 6.0;
            N:=1;
            repeat
                N := N+1;
                C := C * (2*N*(2*N+1)/(2*N-1));
                D := A*(-1)**N * X**(2*N+1)/C;
                B := B+D;
            until (N=71) or (abs(D)<1d-7);
            NORMAL := B;
            end;
        end; {function normal}


Function NORMAL_INVERSA(PROB:double):double;
        {Dada a probabilidade, calcula o valor que a origina
         Utiliza a função Normal}
    Const
        PRECISAO = 1d-10;
    Var
        Z,ZMAX,ZMIN,Y : double;
    Begin
    if (PROB=0) then Z:=-4.5 else if (PROB=1)
                then Z:=4.5
                else
            begin
            Zmin:= -4.5; Zmax:=  4.5;
            repeat
                Z:= 0.5*(Zmin+Zmax);
                Y:= normal(Z);
                if (Y<PROB) then Zmin:= Z else Zmax:= Z;
            until ((Zmax-Zmin)<PRECISAO);
            end;
        NORMAL_INVERSA:= Z;
        end;{function Normal_inversa}


{############## BLOCO DOS PROCEDIMENTOS AUXILIARES ########}


Procedure LE_AMOSTRA (var OBS:vector1; var M:integer);
    var
        F : text; NOME : packed array [1..60] of char;
    Begin
    writeln;
    write('Qual o nome do ficheiro que contém a amostra inicial ? ');
    readln(NOME);
    writeln;
    open(F,NOME,old);
    reset(F);
    M:=0;
    repeat
        M:=M+1;
        readln(F,OBS[m]);
```

```
    until eof(F);
    End; {procedure Le_amostra}


Procedure ESCREVE(MEDIA,VARC,DPC,DIF_MEDIA,M2,M3,M4,G1,G2:double;
                  DIMENSAO:integer);
    Begin
    writeln(SAIDA); writeln(SAIDA); writeln(SAIDA);
    writeln(SAIDA,'     1. Estatísticas iniciais');
    writeln(SAIDA,'        ----------------------');
    writeln(SAIDA);
    writeln(SAIDA,'        Dimensão ..................... ',DIMENSAO);
    writeln(SAIDA,'        Média ........................ ',MEDIA);
    writeln(SAIDA,'        Variância corrigida .......... ',VARC);
    writeln(SAIDA,'        Desvio-padrão corrigido ...... ',DPC);
    writeln(SAIDA,'        Diferença média .............. ',DIF_MEDIA);
    writeln(SAIDA);
    writeln(SAIDA,'        Momento central de 2ª ordem ... ',M2);
    writeln(SAIDA,'        Momento central de 3ª ordem ... ',M3);
    writeln(SAIDA,'        Momento central de 4ª ordem ... ',M4);
    writeln(SAIDA);
    writeln(SAIDA,'        Coeficiente de assimetria ..... ',G1);
    writeln(SAIDA,'        Kurtosis ..................... ',G2);
    writeln(SAIDA); writeln(SAIDA); writeln(SAIDA);
    End; {procedure Escreve}


Procedure CALC_ESTATISTICAS (var MEDIA,VARC,DPC,DIF_MEDIA,M2,M3,M4,G1,
                                 G2:double; X:vector1; MM:integer);
        {As estatísticas são referentes às MM primeiras observações
         contidas no vector X}
    Var
        I,J : integer;
    Begin
    MEDIA:=0.0; DIF_MEDIA:=0.0; M3:=0.0; M4:=0.0; VARC:=0.0;
    for I:=1 to MM do
        begin
        MEDIA:=MEDIA+X[i]; VARC:=VARC+X[i]*X[i];
        for J:=1 to MM do DIF_MEDIA:=DIF_MEDIA+abs(X[i]-X[j]);
        end;
    MEDIA:=MEDIA/MM; M2:=(VARC/MM-sqr(MEDIA)); VARC:=(mm/(MM-1))*M2;
    DPC:=sqrt(VARC); DIF_MEDIA:=DIF_MEDIA/(MM*(MM-1));
    for I:=1 to MM do begin
                M3 := M3+(X[i]-MEDIA)**3; M4 := M4+sqr(sqr(X[i]-MEDIA));
                   end;
    M3:=M3/MM; M4:=M4/MM; G1:=M3/exp(1.5*ln(M2)); G2:=M4/sqr(M2)-3;
    End; {procedure Calcula _estatisticas}

{############## BLOCO DOS PROCEDIMENTOS DE NíVEL 1 ########}


Procedure ESTUDA_AMOSTRA1 (OBS:vector1; MM:integer);
        {Calcula e imprime um conjunto de estatísticas sobre uma
         amostra constituída pelos primeiros MM elementos do vector OBS.
         Recorre aos procedimentos CALC_ESTATíSTICAS e ESCREVE}
   Var
        M2,M3,M4,G1,G2,MEDIA,VARC,DPC,DIF_MEDIA : double;
    Begin
    calc_estatisticas(MEDIA,VARC,DPC,DIF_MEDIA,M2,M3,M4,G1,G2,OBS,MM);
    writeln(SAIDA); writeln(SAIDA); writeln(SAIDA);
    writeln(SAIDA,'      ESTUDO DA AMOSTRA INICIAL');
    escreve(MEDIA,VARC,DPC,DIF_MEDIA,M2,M3,M4,G1,G2,MM);
    End; {procedure Estuda_amostra1}
```

```
Procedure ESTUDA_AMOSTRA3 (var X:vector1; MM:integer);

      {Numa primeira fase, calcula e imprime um conjunto de estatísticas
       sobre uma amostra constituída pelos primeiros MM elementos do
       vector X. Neste ponto, o procedimento assemelha-se a ESTUDA_AMOSTRA1,
       tendo apenas como particularidade o impôr que o output seja escrito
       numa página nova.
       Numa segunda fase, ordena-se a amostra por ordem crescente (vai ser
       reenviada ordenada) para calcular e imprimir algumas estatísticas
       de ordem. Este procedimento também cria o ficheiro GRAFICO2.DAT.
       Recorre-se aos procedimentos CALC_ESTATíSTICAS, ESCREVE e ORDENA}
   Var
       M2,M3,M4,G1,G2,MEDIA,VARC,DPC,DIF_MEDIA : double;
       F                                        : text;
       I                                        : integer;
   Begin
   calc_estatisticas(MEDIA,VARC,DPC,DIF_MEDIA,M2,M3,M4,G1,G2,X,MM);
   writeln(SAIDA,''(12)); writeln(SAIDA);
   writeln(SAIDA); writeln(SAIDA); writeln(SAIDA);
   writeln(SAIDA,'     ESTUDO DAS AMOSTRAS BOOTSTRAP');
   escreve(MEDIA,VARC,DPC,DIF_MEDIA,M2,M3,M4,G1,G2,MM);
   ordena(X,MM);
   open(F,'GRAFICO2.DAT',new);
   rewrite(F);
   for I:=1 to MM do
       begin
       writeln(F,X[i],'  ',dble(I-1)/dble(MM));
       writeln(F,X[i],'  ',dble(I)/dble(MM));
       end;
   close(F);
   writeln(SAIDA,'     2. Estatísticas de ordem');
   writeln(SAIDA,'        ----------------------');
   writeln(SAIDA);
   writeln(SAIDA,'        Percentil a  2.5% ..... ',
       TETA_[trunc(MM*0.025)+1]);
   writeln(SAIDA,'        Percentil a  5.0% ..... ',
       TETA_[trunc(MM*0.050)+1]);
   writeln(SAIDA,'        Mediana           ..... ',
       TETA_[trunc(MM*0.500)+1]);
   writeln(SAIDA,'        Percentil a 95.0% ..... ',
       TETA_[trunc(MM*0.950)+1]);
   writeln(SAIDA,'        Percentil a 97.5% ..... ',
       TETA_[trunc(MM*0.975)+1]);
   writeln(SAIDA);
   writeln(SAIDA,'        A função de distribuição amostral de TETA',
       ' encontra-se no');
   writeln(SAIDA,'        ficheiro GRAFICO2.DAT');
   writeln(SAIDA); writeln(SAIDA); writeln(SAIDA);
   End; {procedure estuda_amostra3}
Procedure ESTUDA_AMOSTRA2(var A:double; OBS:vector1; TETA:double;
                          MM:integer);
      {Numa primeira fase, constrói um vector VEC_TETA, cujo elemento i
       vai ser a estatística TETA calculada sem se incluir na amostra o
       elemento i desta ***     sub-procedimento CONSTROI_VEC_TETA que
       também vai gravar o ficheiro VEC_TETA.DAT.
       Numa segunda fase, calculam-se as estatísticas A e A', bem como o
       SIGMA (Jack) e o BIAS *** sub-procedimentos CALC_A e CALC_ALINHA.
       Numa terceira fase, imprimem-se os resultados ** sub-procedimento
       IMPRIME.
       Recorre-se ao procedimento CALC_TETA.
       Ter em atenção que só a estatística A é enviada para o programa
       principal e que o procedimento CALC_A grava o ficheiro
       GRAFICO1.DAT}
   Var
       MED,ALINHA,SIGJACK,BIAS,TETATIL : double;
       I                               : integer;
```

```
    VEC_TETA                              : vector1;
procedure CONSTROI_VEC_TETA(var X:vector1; var MED:double;
                            Y : vector1; MM:integer);
    var
        I : integer;  AUX : double;  F : text;
    begin
    X[mm] := calc_teta(Y,MM-1);
    for I:=1 to MM-1 do begin
                        swap(Y[i],Y[mm]);
                        X[i] := calc_teta(Y,MM-1);
                        end;
    MED := 0.0;
    for I:=1 to MM do MED := MED + X[i];
    MED := MED/MM;
    open(F,'VEC_teta.DAT',new);
    rewrite(F);
    for I:=1 to MM do writeln(F,X[i]);
    close(F);
    end; {procedure constroi_vec_teta}

procedure CALC_A (var A:double; X,OBS:vector1; MED:double;
                  MM:integer);
    var
        I : integer;  TOT1,TOT2,AUX : double; F : text;
    begin
    open(F,'GRAFICO1.DAT',new);
    rewrite(F);
    TOT1 := 0.0;  TOT2 := 0.0;
    for I:=1 to MM do begin
                      AUX := MED-X[i];
                      TOT1 := TOT1 + sqr(AUX);
                      TOT2 := TOT2 + AUX * sqr(AUX);
                      writeln(F,OBS[i],'  ',(MM-1)*AUX);
                      end;
    writeln(F,OBS[1],'  ',dble(0));
    writeln(F,OBS[mm],'  ',dble(0));
    close(F);
    A := (1.0/6.0)*(TOT2/exp(1.5*ln(TOT1)));
    end; {procedure calc_a}

procedure CALC_ALINHA (var ALINHA:double; X:vector1; MED:double;
                  MM:integer);
    var
        I : integer;  TOT1,TOT2,AUX : double;
    begin
    TOT1 := 0.0;  TOT2 := 0.0;
    for I:=1 to MM do begin
                      AUX := MED-X[i];
                      TOT1 := TOT1 + sqr(AUX);
                      TOT2 := TOT2 + AUX * sqr(AUX);
                      end;
    ALINHA := (1.0/6.0)*(TOT2/exp(1.5*ln(TOT1)));
    end; {procedure calc_alinha}

procedure IMPRIME(MED,A,ALINHA,SIGJACK,BIAS,TETATIL:double);
    begin
    writeln(SAIDA,'     3. Estudo das sub-amostras');
    writeln(SAIDA,'        ----------------------');
    writeln(SAIDA);
    writeln(SAIDA,'        Média do vector TETA ..... ',MED);
    writeln(SAIDA,'        Estatística a ............ ',A);
    writeln(SAIDA,
    '        Estatística a',''(39),' ............... ',ALINHA);
    writeln(SAIDA,'        SIGMA (Jack) ............. ',SIGJACK
    writeln(SAIDA);
    writeln(SAIDA,'        Os elementos do vector TETA',
```

```
        ' encontram-se no ficheiro TETA.DAT');
        writeln(SAIDA,'        e as coordenadas dos pontos (Xi,Ui) no',
        ' ficheiro GRAFICO1.DAT');
        writeln(SAIDA); writeln(SAIDA); writeln(SAIDA);
        writeln(SAIDA,'     4. Ligação entre amostra e sub-amostras');
        writeln(SAIDA,'        ------------------------------------');
        writeln(SAIDA);
        writeln(SAIDA,'        BIAS ..........................  ',BIAS);
        writeln(SAIDA,'        TETA_til ......................  ',TETATIL);
        writeln(SAIDA); writeln(SAIDA); writeln(SAIDA);
        end; {procedure imprime}

    Begin
    constroi_vec_teta(VEC_TETA,MED,OBS,MM);
    calc_a(A,VEC_TETA,OBS,MED,MM);
    calc_alinha(ALINHA,VEC_TETA,TETA,MM);
    SIGJACK := 0.0;
    for I:=1 to MM do SIGJACK := SIGJACK + sqr(MED-VEC_TETA[i]);
    SIGJACK := ((MM-1)/MM) * sqrt(SIGJACK);
    BIAS    := (MM-1)*(MED-TETA); TETATIL := TETA - BIAS;
    imprime(MED,A,ALINHA,SIGJACK,BIAS,TETATIL);
    End; {procedure estuda_amostra2}

Procedure CONSTROI_AMOSTRA (OBS:vector1; MM:integer; var AMOSTRA:vector1;
                            var CONTADOR:vector2; var SEED:unsigned);
        {Constrói, por tiragem aleatória com reposição, amostras de dimensão
         MM (vector AMOSTRA) a partir de um universo de dimensão MM (OBS).
         A geração dos números aleatórios inteiros é feita na função
         RNDINT (ver comentário na função), servindo o vector CONTADOR
         para testes posteriores aos números aleatórios.}
    Var
        I,K : integer;
    function RNDINT(var SEED:unsigned):integer;
        {Gera um número pseudo-aleatório inteiro entre MINIMO e MAXIMO, a
         partir da geração de um real entre 0 e 1, em dupla precisão, pelo
         método linear congruente, isto é,
                SEED (nova) <- [ (SEED(antiga)*69069) + 1] mod 2^32}
        const
            MINIMO = 1;
            B = 4294967296d0; {2**32}
            C = 69069d0;
        var
            A : double;  MAXIMO : integer;
        begin
        MAXIMO := MM;
        A    := C*dble(SEED)+1d0;                {A = (seed*69069) + 1}
        SEED := utrunc(A-B*dble(utrunc(A/B))); {seed = A mod B}
        RNDINT:= trunc((dble(SEED)/B)*dble(MAXIMO)) + MINIMO;
        end; {function RNDINT}

    Begin
    for I:=1 to MM do begin
                      K := rndint(SEED);
                      AMOSTRA[i] := OBS[k];
                      CONTADOR[k] := CONTADOR[k]+1;
                      end;
    End; {procedure constroi_amostra}


Function G(S:double; X:vector1; N:integer):double;
        {Função de distribuição empírica de TETA. G(s)=Prob(S<s).
         O vector X comtém os N valores de TETA observados ordenados
         por ordem crescente}
    var
        I : integer;
    Begin
```

```
    if (S>X[n]) then G:=1
                else begin
                     I:=1;
                     while(S>X[i]) do I := I+1;
                     G := (I-1)/N;
                     end;
    End;


Procedure histograma(X:vector1; MM:integer);
        {Constrói um histograma dos valores de TETA observados no
         bootstrap e imprime o ficheiro HISTO.DAT}
    var
        ANT,I : integer;  INF,SUP : double; F:text;
    Begin
    open(F,'HISTO.DAT',new);
    rewrite(F);
    writeln(F,'          Histograma dos valores de TETA do bootstrap');
    writeln(F,'       INF       SUP     Freq. Absoluta');
    writeln(F,'     ------    ------    --------------');
    I:=1; INF:=trunc(X[1]*100)/100.0;
    repeat
        ANT:=I; SUP:=INF+0.02;
        while( (X[i]<SUP) and (I<MM) ) do I:=I+1;
        if (I=MM) then begin SUP:=X[i]; I:=I+1; end;
        writeln(F,INF:10:2,SUP:10:2,I-ANT);
        INF:=SUP;
    until (I>MM);
    close(F);
    End;


procedure imprime_resultados_b1(A,TETA:double; N:integer;var TETA_:vector1);
      var Z0,Z1,Z2,Z3,Z4,Z5,Z6,Z7,Z8,JE1,JE2,JD1,JD2,IE1,IE2,ID1,ID2 : double;
    begin
    Z0 := normal_inversa(g(TETA,TETA_,N));
    Z1:=2*Z0-1.6449; Z2:=2*Z0+1.6449; Z3:=2*Z0-1.96; Z4:=2*Z0+1.96;
    Z5:=Z0+((Z0-1.6449)/(1-A*(Z0-1.6449)));
    Z6:=Z0+((Z0+1.6449)/(1-A*(Z0+1.6449)));
    Z7:=Z0+((Z0-1.96)/(1-A*(Z0-1.96)));
    Z8:=Z0+((Z0+1.96)/(1-A*(Z0+1.96)));
    writeln(SAIDA,'     3. Valores auxiliares');
    writeln(SAIDA,'        -------------------');
    writeln(SAIDA);
    writeln(SAIDA,'        G(TETA) ..................... ',G(TETA,TETA_,N));
    writeln(SAIDA,'        Z0 .......................... ',Z0);
    writeln(SAIDA,'        Z1 e Z2 ..................... ',Z1,'  ',Z2);
    writeln(SAIDA,'        Z3 e Z4 ..................... ',Z3,'  ',Z4);
    writeln(SAIDA,'        Z5 e Z6 ..................... ',Z5,'  ',Z6);
    writeln(SAIDA,'        Z7 e Z8 ..................... ',Z7,'  ',Z8);
    writeln(SAIDA); writeln(SAIDA); writeln(SAIDA);
    IE1 := TETA_[trunc(N*normal(Z1))+1]; ID1 := TETA_[trunc(N*normal(Z2))+1];
    IE2 := TETA_[trunc(N*normal(Z3))+1]; ID2 := TETA_[trunc(N*normal(Z4))+1];
    JE1 := TETA_[trunc(N*normal(Z5))+1]; JD1 := TETA_[trunc(N*normal(Z6))+1];
    JE2 := TETA_[trunc(N*normal(Z7))+1]; JD2 := TETA_[trunc(N*normal(Z8))+1];
    writeln(SAIDA,'     4. Intervalos de confiança');
    writeln(SAIDA,'        ------------------------');
    writeln(SAIDA);
    writeln(SAIDA,'        Não corrigido a 90% ......... ',IE1,'  ',ID1);
    writeln(SAIDA,'        Não corrigido a 95% ......... ',IE2,'  ',ID2);
    writeln(SAIDA,'        Corrigido a 90% ............. ',JE1,'  ',JD1);
    writeln(SAIDA,'        Corrigido a 95% ............. ',JE2,'  ',JD2);
    writeln(SAIDA); writeln(SAIDA); writeln(SAIDA);
    end;{procedure imprime_resultados_b1}
```

```
procedure LE_ELEMENTOS (var N1,N2: integer; var SEED: unsigned);
        var R:char;
    begin
    writeln;
    repeat
        write('Quantas amostras para o "bootstrap" principal ? ');
        readln(N1);
        if (N1>M_N_MAX) then writeln ('O número máximo previsto é',M_N_MAX);
    until (N1>0) and (N1<=M_N_MAX);
    repeat
        write('Vai efectuar um "bootstrap" iterativo (s/n) ? ');
        readln(R);
    until (R='S') or (R='s') or (R='N') or (R='n');
    if (R='S') or (R='s') then
        repeat
            write('Quantas amostras para o "bootstrap" iterativo ? ');
            readln(N2);
            if (N2>M_N_MAX) then writeln ('O número máximo previsto é',M_N_MA
        until (N2>0) and (N1<=M_N_MAX)
                            else N2:=0;
    write('Qual o gerador inicial ? ');
    readln(SEED);
    writeln;
    end;{procedure LE_ELEMENTOS}


procedure BOOTSTRAP_2(M,NUMERO,LINHA:integer; TETA:double; var SEED:unsigned;
                      var BASE:vector1; var MAT_TETA:matriz);
        var T1,T2:double; I: integer; CONTADOR:vector2; AMOSTRA1:vector1;
    begin
    T1:=0; T2:=0;
    for I:=1 to NUMERO do
        begin
        constroi_amostra(BASE,M,AMOSTRA1,CONTADOR,SEED);
        MAT_TETA[linha,i]:=calc_teta(AMOSTRA1,M);
        T1:=T1+MAT_TETA[linha,i]; T2:=T2+MAT_TETA[linha,i]*MAT_TETA[linha,i];
        end;
        T1:=T1/NUMERO; T2:=T2/NUMERO-T1*T1; T2:=sqrt(T2*NUMERO/(NUMERO-1));
    for I:=1 to NUMERO do
        MAT_TETA[linha,i]:=(MAT_TETA[linha,i]-TETA)/T2;
    ordena1(MAT_TETA,LINHA,NUMERO);
    end;


procedure conclui_bootstrap_2(var TETA_:vector1; var MAT_TETA:matriz;
                        TETA,T2:double; N1,N2:integer);
        VAR I,C,R1,R2,R3,R4: integer; AUX,L1,L2,L3,L4:double;
            TETA1_: vector1; F:text;

    begin
    for I:=1 to N1 do
        begin
        AUX:=(TETA_[i] - TETA)/T2;
        if (MAT_TETA[i,n2]<=AUX) then C:=N2
                                 else
            begin
            C:=1;
            while (MAT_TETA[i,c]<=AUX) and (C<N2) do C:=C+1;
            C:=C-1;
            end;
        TETA1_[i]:=C/N2;
        TETA_[i] :=AUX;
        end;
    ordena(TETA_,N1); ordena(TETA1_,N1);
    open(F,'VECTORES.DAT',new); rewrite(F);
    for I:=1 to N1 do writeln(F,TETA_[i],'   ',TETA1_[i]);
```

```
    close(F);
    L1:=TETA1_[trunc(N1*0.025+0.99)]; L2:=TETA1_[trunc(N1*0.05+0.99)];
    L3:=TETA1_[trunc(N1*0.95+0.99)];  L4:=TETA1_[trunc(N1*0.975+0.99)];
    writeln(SAIDA,'      6. Resultados preliminares do bootstrap iterativo');
    writeln(SAIDA,'         ----------------------------------------------');
    writeln(SAIDA);
    writeln(SAIDA,'          L1 ........ ',L1);
    writeln(SAIDA,'          L2 ........ ',L2);
    writeln(SAIDA,'          L3 ........ ',L3);
    writeln(SAIDA,'          L4 ........ ',L4);
    R1:=max(trunc(N1*L1+0.99),1); R2:=max(trunc(N1*L2+0.99),1);
    R3:=trunc(N1*L3+0.99); R4:=trunc(N1*L4+0.99);
    writeln(SAIDA);
    writeln(SAIDA,'          R1 ........ ',R1);
    writeln(SAIDA,'          R2 ........ ',R2);
    writeln(SAIDA,'          R3 ........ ',R3);
    writeln(SAIDA,'          R4 ........ ',R4);
    writeln(SAIDA); writeln(SAIDA); writeln(SAIDA);
    L1:=TETA_[r1]; L2:=TETA_[r2]; L3:=TETA_[r3]; L4:=TETA_[r4];
    writeln(SAIDA,'      7. Resultados do bootstrap iterativo');
    writeln(SAIDA,'         ---------------------------------');
    writeln(SAIDA);
    writeln(SAIDA,'        Número de amostras por iteração ... ',N2);
    writeln(SAIDA);
    writeln(SAIDA,'          Intervalo a 95% ..... ',
        TETA-L4*T2,'    ',TETA-L1*T2);
    writeln(SAIDA,'          Intervalo a 90% ..... ',
        TETA-L3*T2,'    ',TETA-L2*T2);
    writeln(SAIDA);
    writeln(SAIDA,
        '         As distribuições base encontram-se no ficheiro VECTORES.DAT');
    writeln(SAIDA); writeln(SAIDA);
    end;

{############## BLOCO DE TESTE AOS NÚMEROS ALEATÓRIOS  ##########}

Procedure testa_numeros(contador:vector2;M,N:integer);
        {Imprime a frequência de saída de cada um dos M números bem
         como a frequência em cada classe de probabilidade 5% dada
         pelo T.L.Central}
    Var
        DP : double;  SAI : boolean;   I,J : integer;   F : text;
        LIMITE : array [1..19] of integer;   VINTIL : vector2;
    Begin
    DP := sqrt(dble(N*(M-1))/dble(M)); LIMITE[10] := N; VINTIL := zero;
    LIMITE[1] := trunc(N - 1.645*DP);  LIMITE[19] := trunc(N + 1.645*DP);
    LIMITE[2] := trunc(N - 1.282*DP);  LIMITE[18] := trunc(N + 1.282*DP);
    LIMITE[3] := trunc(N - 1.036*DP);  LIMITE[17] := trunc(N + 1.036*DP);
    LIMITE[4] := trunc(N - 0.842*DP);  LIMITE[16] := trunc(N + 0.842*DP);
    LIMITE[5] := trunc(N - 0.674*DP);  LIMITE[15] := trunc(N + 0.674*DP);
    LIMITE[6] := trunc(N - 0.524*DP);  LIMITE[14] := trunc(N + 0.524*DP);
    LIMITE[7] := trunc(N - 0.385*DP);  LIMITE[13] := trunc(N + 0.385*DP);
    LIMITE[8] := trunc(N - 0.253*DP);  LIMITE[12] := trunc(N + 0.253*DP);
    LIMITE[9] := trunc(N - 0.126*DP);  LIMITE[11] := trunc(N + 0.126*DP);
    for I:=1 to M do
        begin
        J := 0;   SAI := false;
        repeat
            J:=J+1;
            if (CONTADOR[i]<=LIMITE[j]) then begin
                                                VINTIL[j] := VINTIL[j] + 1;
                                                SAI := true;
                                                end;
        until SAI or (J=19);
        if not SAI then VINTIL[20] := VINTIL[20] + 1;
        end;
```

```
    open(F,'ALEATORIO.DAT',new);
    rewrite(F);
    writeln(F,'         Impressão das frequências de saída das',
              ' diferentes observações');
    writeln(F);
    I:=0;
    repeat
        J:=0;
        repeat
            J:=J+1; I:=I+1;
            write(F,contador[i]);
        until (J=8) or (I=M);
        writeln(F);
    until(I=M);
    writeln(F); writeln(F);
    writeln(F,'Número de ocorrências dentro dos intervalos de 5 em 5%',
              ' dados pelo T. L. Central');
    writeln(F);
    for J:=0 to 1 do
        begin
        for I:=1 to 9 do write(F,VINTIL[10*j+i]:8);
        writeln(F,VINTIL[10*j+10]:8);
        end;
    close(F);
    End; {procedure testa_numeros}


{************** PROGRAMA PRINCIPAL ********}


BEGIN
{PARTE 1  ------------------------------------------------------}

{Fase 1 -- Abertura do ficheiro de output, leitura e ordenação da amostra}
open(SAIDA,'RES.DAT',new);
rewrite(SAIDA);
le_amostra(OBS,M);
ordena(OBS,M);

{Fase 2 -- Estudo da amostra}
estuda_amostra1(OBS,M);

{Fase 3 -- Calculo da estatística que vai ser objecto do "bootstrap" (TETA)
           e saída para output}
nome(NOME_TETA);
TETA := calc_teta(OBS,M);
writeln(SAIDA,'      2. Estatística em Estudo');
writeln(SAIDA,'         ----------------------');
writeln(SAIDA);
writeln(SAIDA,'         ',NOME_TETA,'(TETA) ..... ',TETA);
writeln(SAIDA); writeln(SAIDA); writeln(SAIDA);

{Fase 4 -- Construção e estudo das M sub-amostras de dimensão (M-1)}
estuda_amostra2(A,OBS,TETA,M);


{PARTE 2 ------------------------------------------------------}

{Fase 1 -- Geração das N amostras bootstrap e cálculo do vector TETA_}
CONTADOR := zero; T1:=0; T2:=0;
le_elementos(N1,N2,SEED);
for I:=1 to N1 do
    begin
    constroi_amostra(OBS,M,AMOSTRA,CONTADOR,SEED);
    TETA_[i] := calc_teta(AMOSTRA,M);
```

```
    T1:=T1+TETA_[i]; T2:=T2+TETA_[i]*TETA_[i];
    bootstrap_2(M,N2,I,TETA_[i],SEED,AMOSTRA,MAT_TETA);
    end;
T1:=T1/N1; T2:=(T2/N1)-(T1*T1); T2:=sqrt(N1*T2/(N1-1));
if NOME_TETA='Índice de Gini' then histograma(TETA_,N1);

{Fase 2 -- Estudo dos resultados do bootstrap}
{etapa 1 ... Bootstrap principal ... (tipo 2 e 3)  }
estuda_amostra3(TETA_,N1);
imprime_resultados_b1(A,TETA,N1,TETA_);
{etapa 2 ... Bootstrap principal ... (tipo 4) }
writeln(SAIDA,chr(12));
writeln(SAIDA,'      5. Bootstrap T');
writeln(SAIDA,'         -----------');
writeln(SAIDA);
writeln(SAIDA);
writeln(SAIDA,'         Intervalo a 95% ..... ',
        2*TETA-TETA_[trunc(N1*0.975)+1],'    ',2*TETA-TETA_[trunc(N1*0.025)+1]
writeln(SAIDA,'         Intervalo a 90% ..... ',
        2*TETA-TETA_[trunc(N1*0.950)+1],'    ',2*TETA-TETA_[trunc(N1*0.050)+1]
writeln(SAIDA); writeln(SAIDA); writeln(SAIDA);
{etapa 3 ... Bootstrap secundário ... }
conclui_bootstrap_2(TETA_,MAT_TETA,TETA,T2,N1,N2);

close(SAIDA);


{PARTE 3 ------------------------------------------------------}
testa_numeros(CONTADOR,M,N1);


END.
```

7 - BIBLIOGRAFIA

ABRAMOVITCH, L.; **SINGH**, K. (1985)
"Edgeworth corrected pivotal statistics and the Bootstrap"
The Annals of Statistics, 13, 1, 116-132

ARVENSEN, J. N.; **SALSBURG**, D. S. (1973)
"Approximate tests and confidence intervals using the Jackknife"
Perspectives in Biometrics, Editado por R. M. Elashoff, Academic Press, New York

BABU, G. J.; **BOSE**, A. (1989)
"Bootstrap confidence intervals"
Statistics & Probability Letters, 7, 2, 151-160

BERAN, R. (1982)
"Estimated sampling distributions: the Bootstrap and competitors"
The Annals of Statistics, 10, 1, 212-225

BERAN, R. (1984)
"Jackknife approximations to Bootstrap estimates"
The Annals of Statistics, 12, 1, 101-118

BERAN, R. (1987)
"Prepivoting to reduce level error of confidence sets"
Biometrika, 74, 3, 457-468

BERAN, R. (1988)
"Prepivoting test statistics: a Bootstrap view of asymptotic refinements"
Journal of the American Statistical Association, 83, 403, 687-697

BERAN, R.; **SRIVASTAVA**, M. S. (1985)
"Bootstrap tests and confidence regions for functions of a covariance matrix"
The Annals of Statistics, 13, 1, 95-115
NOTA: posteriormente foi realizada, pelos autores, uma correcção a este artigo, a qual foi editada em The Annals of Statistics, 1987, 15, 1, 470-471

BICKEL, P. J.; **FREEDMAN**, D. A. (1981)
"Some asymptotic theory for the Bootstrap"
The Annals of Statistics, 9, 6, 1196-1217

BICKEL, P. J.; **KRIEGER**, A. M. (1989)
"Confidence bands for a distribution function using the Bootstrap"
Journal of the American Statistical Association, 84, 405, 95-100

BUCKLAND, S. T. (1983)
"Monte Carlo methods for confidence intervals using Bootstrap technique"
Bias, 10, 194-212

DAVISON, A. C.; **HINKLEY**, D. V.; **SCHECHTMAN**, E. (1986)
"Efficient Bootstrap simulation"
Biometrika, 73, 3, 555-566

DICICIO, T. J.; **ROMANO**, J. P. (1988)
"A review of Bootstrap confidence intervals"
Journal of the Royal Statistical Society, B, 50, 3, 338-370

DICICIO, T. J.; **TIBSHIRANI**, R. (1987)
"Bootstrap confidence intervals and Bootstrap approximations"
Journal of the American Statistical Association, 82, 397, 163-170

EFRON, B. (1979)
"Bootstrap methods: another look at the Jackknife"
The Annals of Statistics, 7, 1, 1-26

EFRON, B. (1980)
"Computer intensive methods in statistics"
Some Recente Advances in Statistics, Editado por J. Tiago de Oliveira e Benjamin Epstein, Academia das Ciências de Lisboa, Lisboa

EFRON, B. (1981)
"Nonparametric standard errors and confidence intervals"
The Canadian Journal of Statistics, 9, 2, 139-172

EFRON, B. (1982a)
The Jackknife, the Bootstrap and Other Resampling Plans
Society for Industrial and Applied Mathematics, Philadelphia

EFRON, B. (1982b)
"Transformation theory: how normal is a family of distributions?"
The Annals of Statistics, 10, 2, 323-339 e 1032

EFRON, B. (1985)
"Bootstrap confidence intervals for a class of parametric problems"
Biometrika, 72, 1, 45-58

EFRON, B. (1987)
"Better Bootstrap confidence intervals"
Journal of the American Statistical Association, 82, 397, 171-200

EFRON, B.; **GONG**, G. (1983)
"A leisurely look at the Bootstrap, the Jackknife, and cross-validation"
The American Statistician, 37, 1, 36-48

EFRON, B.; **TIBSHIRANI**, R. (1986)
"Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy"
Statistical Science, 1, 1, 54-77

FALK, M. (1986)
"On the accuracy of the Bootstrap approximation of the joint distribution of sample quantiles"
Commercial Statistical Association, 15, 2867-2876

FLOOD, L. (1985)
"Using Bootstrap to obtain standard errors of system Tobit coefficients"
Economics Letters, 19, 339-342

FREEDMAN, D. A. (1984)
"On Bootstrapping two-stage least-squares estimates in stationary linear models"
The Annals of Statistics, 12, 3, 827-842

HALL, P. (1983)
"Inverting an Edgeworth Expansion"
The Annals of Statistics, 11, 2, 569-576

HALL, P. (1986a)
"On the Bootstrap and confidence intervals"
The Annals of Statistics, 14, 4, 1431-1452

HALL, P. (1986b)
"On the number of Bootstrap simulations required to construct a confidence interval"
The Annals of Statistics, 14, 4, 1453-1462

HALL, P. (1987a)
"On the Bootstrap and continuity correction"
Journal of the Royal Statistical Society, B, 49, 1, 82-89

HALL, P. (1987b)
"On the Bootstrap and likelihood-based confidence regions"
Biometrika, 74, 3, 481-493

HALL, P. (1988a)
"On symmetric Bootstrap confidence intervals"
Journal of the Royal Statistical Society, B, 50, 1, 35-45

HALL, P. (1988b)
"Theoretical comparison of Bootstrap confidence intervals"
The Annals of Statistics, 16, 3, 927-985

HALL, P. (1989)
"Antithetic resampling for the Bootstrap"
Biometrika, 76, 4, 713-724

HALL, P.; **MARTIN**, M. A. (1988)
"On Bootstrap resampling and iteration"
Biometrika, 75, 4, 661-671

HALL, P.; **THOMAS**, J.; **DICICIO**, T. J.; **ROMANO**, J. P. (1989)
"On smoothing and the Bootstrap"
The Annals of Statistics, 17, 2, 692-704

HÄRDLE, W.; **BOWMAN**, A. W. (1988)
"Bootstrapping in nonparametric regression: local adaptative smoothing and confidence bands"
Journal of the American Statistical Association, 83, 401, 102-110

HINKLEY, D. V. (1977)
"Jackknife confidence limits using Student-t approximations"
Biometrika, 64, 1, 21-28

HINKLEY, D. V. (1988)
"Bootstrap methods"
Journal of the Royal Statistical Society, B, 50, 3, 321-337 e 355-370

HINKLEY, D. V.; **WEI**, B. C. (1984)
"Improvements of Jackknife confidence limit methods"
Biometrika, 71, 2, 331-339

INSTITUTO NACIONAL DE ESTATÍSTICA (1985)
Inquérito às Receitas e Despesas Familiares 1980/81
Imprensa Nacional-Casa da Moeda, Lisboa

JOHNS, M. V. (1988)
"Importance sampling for Bootstrap confidence intervals"
Journal of the American Statistical Association, 83, 403, 709-714

LO, A. Y. (1987)
"A large sample study of the bayesian Bootstrap"
The Annals of Statistics, 15, 1, 360-375

LO, A. Y. (1988)
"A bayesian Bootstrap for a finite population"
The Annals of Statistics, 16, 4, 1684-1695

LOH, W. Y. (1987)
"Calibrating confidence coefficients"
Journal of the American Statistical Association, 82, 397, 155-162

MEHRAN, F. (1975)
"Bounds on the Gini Index based on observed points of the Lorenz Curve"
Journal of the American Statistical Association, 70, 349, 64-66

MURTEIRA, B. J. F. (1979)
Probabilidades e Estatística
Volume I, Editora McGraw-Hill de Portugal, Lisboa

MURTEIRA, B. J. F. (1980)
Probabilidades e Estatística
Volume II, Editora McGraw-Hill de Portugal, Lisboa

MURTEIRA, B. J. F. (1988a)
Estatística: Inferência e Decisão
Imprensa Nacional-Casa da Moeda, Lisboa

MURTEIRA, B. J. F. (1988b)
"Intervalos de confiança "Bootstrap". Aplicação ao Índice de Gini"
Documento de trabalho nº 48, Centro de Matemática Aplicada à Previsão e Decisão Económica, Instituto Superior de Economia, Lisboa

RODRIGUES, C. F. (sem data)
Estudo Comparativo da Desigualdade nas Despesas das Famílias Portuguesas [1973/74 - 1980/81]
Centro de Investigação Sobre Economia Portuguesa, Instituto Superior de Economia, Lisboa

RUBIN, D. B. (1981)
"The bayesian Bootstrap"
The Annals of Statistics, 9, 1, 130-134

SCHENKER, N. (1985)
"Qualms about Bootstrap confidence intervals"
Journal of the American Statistical Association, 80, 390, 360-361

SEN, P. K. (1986)
"The Gini coefficient and poverty indexes: some reconciliations"
Journal of the American Statistical Association, 81, 396, 1050-1057

SILVERMAN, B. W.; **YOUNG**, G. A. (1987)
"The Bootstrap: to smooth or not smooth?"
Biometrika, 74, 3, 469-479

SINGH, K. (1981)
"On the asymptotic accuracy of Efron's Bootstrap"
The Annals of Statistics, 9, 6, 1187-1195

WITHERS, C. S. (1983)
"Expansions for the distribution and quantiles of a regular functional of the empirical distribution with applications to nonparametric confidence intervals"
The Annals of Statistics, 11, 2, 577-587

WITHERS, C. S. (1984)
"Asymptotic expansions for distributions and quantiles with power series cumulants"
Journal of the Royal Statistical Society, B, 46, 389-396

WU, C. F. J. (1986)
"Jackknife, Bootstrap and other resampling methods in regression analysis"
The Annals of Statistics, 14, 4, 1261-1350